International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

Conversion System Raw Data into Structured Data in Building the Vietnamese-Ede Bilingual Vocabulary Corpus

Le Hoang Thi My
Danang College of Technology
The University of Danang
Danang, Vietnam

***Abstract*— In the Vietnamese language processing, some corpus have shared for research purposes. Besides, results of ethnic minority corpus in general and Ede language in particular are based on researching on the small corpus, neither inheriting the results of research on Vietnamese corpus nor using Unicode. Therefore, the Ede language processing has not inherited the results of the Ede language corpus. The objective of this study is proposing the conversion system raw data into structured data basing on the Vietnamese-Ede interactive model in building the Vietnamese-Ede vocabulary corpus. This system can be expanded for building the Vietnamese-ethnic minority bilingual vocabulary corpus.**

Keywords: Bilingual vocabulary corpus, Ede language, Interactive model, Conversion system, Natural language processing

I. INTRODUCTION

In natural language processing (NPL) for Vietnamese in general and ethnic minority languages in particular the tools and resources such as dictionaries for machine, corpus are indispensable in NLP [10]. Vocabulary corpus is used to store all vocabulary words (single and compound words), phrases and sentences. Bilingual vocabulary corpus is the basis for building Educational applications, such as bilingual teaching, building a dictionary to search for the word, check spelling of words and translate-automatically bilingual lessons...

So far, the results of research on the Vietnamese-Ede and Ede-Vietnam vocabulary corpus have been announced, such as building of Vietnamese - Ede vocabulary corpus [6], [8]; building Vietnamese-Ede dictionary used to search for words in the translation of bulletin from Vietnamese into Ede of the DakLak Radio and Television station [2]; Ede-Vietnamese dictionary file on Violet website for teaching Ede language [13]. These results have contributed in Ede language processing such as searching meaning of word, translating manually the bulletin from Vietnamese into Ede language of the Radio Voice of Vietnam in DakLak, teaching Ede language. However, so far the research results on the Vietnamese-Ede and Ede-Vietnamese vocabulary corpus still have some disadvantages that have not been overcome, such as:

- Lack of unity in using Unicode fonts (Arial, Times New Roman, Tahoma, Verdana, etc...) in corpus.
- Building the testing model with small dataset.
- No processing the word-boundary ambiguity, polysemous word, homonyms and part of speech.
- Lack of unity among researchers
- Most researches have been unsystematic, sporadic, lack of clear orientation..

From the above situations, the paper presents the Vietnamese-Ede interactive model for developing the Vietnamese-Ede and Ede-Vietnamese vocabulary corpus with the aim of overcoming the real situations of the current Ede language vocabulary corpus and building the infrastructure for Ede language processing.

II. OVERVIEW OF CORPUS

Databank which contains texts, tables, schemas, images, sound, voice... has been long time and is referred to as databases. Interpreted data as language material and is organized into own type called corpus. Corpus with a single or two languages is called monolingual or bilingual one, respectively and corpus with three or more languages is called multilingual one.

Parallel corpus is bilingual and multilingual corpus that are formated to be able to compare side by side. To contribute to researches on natural language processing and especially in building dictionary, corpus is often added the following information: annotating information on part of speech of words in phrase is called part of speech tagging; Separating and tagging phrase in sentence is called phrase segmenting. For isolating languages such as Vietnamese, there are identifying and tagging procedures for word unit due to unclear word boundaries [10].

A. The role of the corpus in NLP

The monolingual and multilingual corpus are always an indispensable resource in every operation relating to NLP. The applications such as text editors, sending message, chat, email, searching for information online, machine translation, text analysis, games, automatic reference... always have the interference with NLP. The difficulty of the NLP problem increases from lexical, syntax, to semantic pragmatic and pragmatic. The common tasks are word segmentation and part-of-speech tagging. That is the determination of the word/phrase which is present in each sentence of the document, part of speech (noun, verb, adjective...), semantic and grammatical functions of them. These operations are easy for human, but very difficult to solve in NLP and the most difficult problem is the ambiguity in natural language.

Most of the above processing are related to corpus. The quality of the bilingual corpus plays a decisive role in the output quality of the translation system such as analysis and synthesis of texts, machine translation.... Especially, the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

statistical machine translation systems will not produce a reasonable output, if the quality of corpus used in the training process is not good, although the most advanced machine learning methods are applied.

B. Data source in corpus

For the monolingual corpus, data is gathered from the following sources:

- E-books: various specialized books.
- Dictionary: each entry has the examples and explaining; the language standard is accurate and contents are very rich.
- Internet: data is huge and stored on computer (not being updated manually). This data source has many various fields and formatting, so that data must be standardized.
- The corpus are built by the linguists and information technology engineers such as PTB (Penn Tree Bank), SUSANNE (Surface and Underlying Structural Analyses of Naturalistic English)... [5].

So far, bilingual corpus has many electronic data sources that have been translated into many languages. However, the translation has the disadvantages such as free translation, translating main idea, no translating 1:1.

III. The Conversion System Raw Data Into Structured Data Basing On The Vietnamese-Ede Interactive Model

A. The Vietnamese-Ede and Ede-Vietnamese vocabulary corpus

1) Using Unicode in the corpus: The Ede letters also uses the Latin characters, with 76 Ede characters (uppercase and lowercase letters) as seen in table I [3].

The Ede letters includes 68 characters of Unicode and 8 characters of non-Unicode (ẽ, õ, ũ, ĩ, Ẽ, Õ, Ũ, Ĩ). So far, there is still no research using Unicode in the Ede corpus. In this paper, the research results on using Unicode in encoding the Ede language [4] are applied in encoding the Ede corpus.

2) Criteria for the corpus: Based on building infrastructure for Ede language processing, we offer the following criteria:

- The Ede words have been collected and written follow Kpă group of Ede language that is comprehensible and clear voice dialect. The Ede entries reflect Ede culture and tradition. Ede language is written in Ede letters.
- The Vietnamese words belong Vietnamese (Kinh) language and are written in National script.
- The examples are added to clarify meaning and context of entry.
- Entry's part of speech is tagged label with character "N" for noun, "V" for verb, "A" for adjective and "O" for other part of speech.
- Polysemy entry is recorded for translating and comparing with different entries in target language.
- While bilingual translating, starting from source language word and finding target language words are equivalent to original and frequently used meaning in both languages.
- Using Unicode for the Ede letters in the corpus.

TABLE I. EDE ALPHABET

Consonant	Uppercase	B	Ɓ	Č	D	Đ	G	H
		J	K	L	M	N	Ñ	P
		R	S	T	W	Y		
	Lowercase	b	ɓ	č	d	đ	g	h
		j	k	l	m	n	ñ	p
		r	s	t	w	y		
Vowel	Uppercase	A	Ă	Â	E	Ĕ	Ê	Ê
		I	Ĭ	O	Ŏ	Ô	Ô	Ô
		Ơ	Ŏ	U	Ŭ	Ư	Ŭ	
	Lowercase	a	ă	â	e	ĕ	ê	ê
		i	ĭ	o	ŏ	ô	ŏ	ŏ
		ơ	ŏ	u	ŭ	ư	ŭ	

Where:

Uppercase

Lowercase

3) The data source: Currently, the Vietnamese-Ede and Ede-Vietnamese electronic bilingual data sources have used own fonts (not using Unicode fonts). For that reason, these electronic data sources can not be inherited for building the vocabulary corpus. The bilingual sources on paper are used in the interactive model:

- The Vietnamese-Ede dictionary includes about 1,000 entries that are basic and commonly used Vietnamese entries [1].
- Ede-Vietnamese dictionary includes about 10,000 entries that are daily used Ede entries [9].

In this paper, the Ede-Vietnamese dictionary is the input of the interactive model for developing the bilingual vocabulary corpus Ede-Vietnamese and Vietnamese-Ede. This dictionary is compiled in 2014 and used in schools, other education institution; college; university; management agencies and cultural activities, news media…

The input data of this model is updated manually based on the Viet-Ede dictionary and saved on the text files with general form.

The Vietnamese monolingual vocabulary corpus of the VLSP theme [12] includes over 31,000 entries and is inherited for inputting in the interactive model.

4) Database of the vocabulary corpus: The relational database model is designed for the Vietnamese- Ede and Ede-Vietnamese bilingual vocabulary corpus. Figure 1 shows structure of the Ede-Vietnamese and Vietnamese-Ede corpus. The relational database is used as a set of tables to store data and entities which are in a relationship. The tables in the relational database is similar to corpus. Structure and data are stored independently. The advantages and disadvantages of the relational database model include:

a) Advantages

+ Separate corpus
+ Independence of data structure
+ Accessing database flexibly
+ Creating suitable applications for user
+ Converting to other formats easily

b) Disadvantages

+ Covering almost completely the physical structure of database. Therefore, it requires the operating system and configuration is strong enough to support for accessing and updating data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

+ This database structure is slightly slower than other databases because of covering almost completely physical structures of database.

However, the number of entries in the Vietnamese-Ede and Ede-Vietnamese corpus are not too much as well as high technology is more and more developing. These disadvantages are also acceptable.

Ede-Vietnamese Vietnamese-Ede

CSO: index; TULOAI: part of speech;
EXPLE: example; BSUNG: supplement

Figure 1. Structured Data of the Ede-Vietnamese and Vietnamese-Ede corpus

B. *The coversion system raw data into structured data*

The conversion in the system is converting the text files of the Vietnamese-Ede and Ede-Vietnamese bilingual dictionaries into develop Vietnamese-Ede and Ede-Viet bilingual vocabulary corpus. Conversion environment includes the functional blocks (1), (2), (3) và (4) in the system as seen in figure 2.

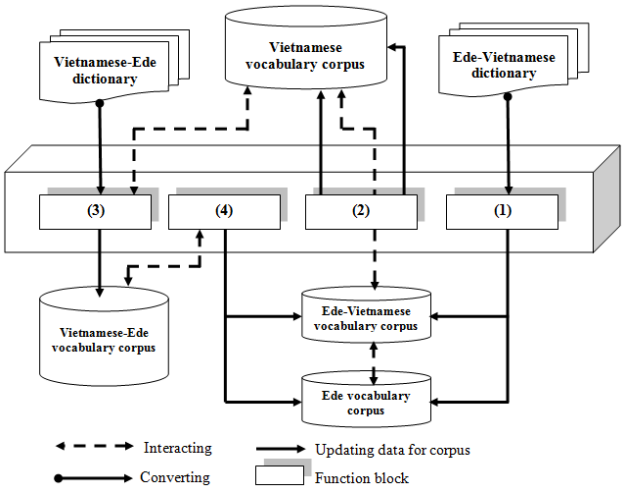

Figure 2. The conversion system

The input of the model: The Vietnamese-Ede and Ede-Vietnamese dictionaries that are updated manually and saved in the text files; the Vietnamese vocabulary Vietnamese corpus is inherited from VLSP website.

The output of the model: the Vietnamese-Ede and Ede-Vietnamese vocabulary corpus, the Ede vocabulary corpus, the Vietnamese corpus with supplemented entries.

The algorithms of the functional blocks

1) The functional blocks (1) and (2): the input is the text files of Ede-Vietnamese dictionary.

a) The functional block (1)

Input: TG(Ede, Viet, Tuloai, Exple) that is imported from the text files of Ede-Vietnamese dictionary.

Output: Ede-Vietnamese vocabulary corpus (B), Ede (D)

Intermedium: B_1: CSO_EDE of B, B_2: VIET attribute of B, B_3: TULOAI attribute of B, D_2: EDE attribute of D, WE: Ede entry, WV: Vietnamese Word, SP: part of speech, CSE: Ede index, $EX_{E\text{-}V}$: Ede-Vietnamese bilingual example.

Method

+ Read(TG; WE, {WV}, SP, {$EX_{E\text{-}V}$})
  - if $WE \in D_2 \rightarrow$ Read(D; CSE)
    - if ($CSE \in B_1$ & $SP \notin B_3$) or ($CSE \in B_1$ & $SP \in B_3$ & $\{WV\} \not\subset B_2$) then Insert(B; CSE, {WV}, SP, {$EX_{E\text{-}V}$})
  - if $WE \notin D_2$
    - Insert(D; WE);
    - Read(D; CSE);
    - Insert(B; CSE, {WV}, SP, {$EX_{E\text{-}V}$})
+ Repeat Read(TG; WE) until EOF(TG)=true

b) The functional block (2): interacting Vietnamese corpus with Ede and Ede-Viet corpus

Input: The Vietnamese (A) and Ede (D) corpus, the Ede-Vietnamese corpus (B)

Output: The Vietnamese-Ede corpus (C), the Vietnamese corpus with supplemented entries.

Intermedium: A_2: VIET attribute of A, C_1: CSO_VIET attribute of C, C_2: EDE attribute of C, C_3: TULOAI attribute of C, CSE: Ede index, CSV: Vietnamese index, WV: Vietnamese Word, EWV: Vietnamese entry, WE_{CSE}: Ede Word aligns with Ede index, SP: part of speech, $EX_{E\text{-}V}$: ví dụ song ngữ Êđê-Việt, $EX_{V\text{-}E}$: Vietnamese-Ede bilingual example, note: noting for supplemented entries

Method

+ Read(B; CSE, {WV}, SP, { $EX_{E\text{-}V}$}), Read(D, WE_{CSE})
+ SplitWord_V {WV} = (EWV_1, EWV_2,…, EWV_n)
+ SpiltExample_E_V {$EX_{E\text{-}V}$} = ({$EX_{E\text{-}V1}$}, {$EX_{E\text{-}V2}$}, …, {$EX_{E\text{-}Vn}$}) on EWV_i
+ In turn EWV_i interacts with (A)
  - If $EWV_i \in A_2$ then Read(A; CSV)
    - if ($CSV \in C_1$ & $SP \notin C_3$) or ($CSV \in C_1$ & $SP \in C_3$ & $WE \notin C_2$) then Insert(C; CSV, WE_{CSE}, SP, ConvertEV({$EX_{E\text{-}Vi}$},{$EX_{V\text{-}Ei}$})
  - $EWV_i \notin A_2$
    - Insert(A; EWV_i, note); Read(A; CSV)
    - Insert(C; CSV, WE_{CSE}, SP, ConvertEV({$EX_{E\text{-}Vi}$},{$EX_{V\text{-}Ei}$})
+ Repeat Read(B; WV) until EOF(B)=true

2) The functional blocks (3) và (4): the output is the text files of Vietnamese-Ede dictionary.

a) The function block (3)

Input: TG(Viet, Ede, Tuloai, Exple) that is imported from the text files of Vietnamese-Ede dictionary.

Output: Vietnamese-Ede vocabulary corpus (C), the Vietnamese corpus with supplemented entries

Intermedium: A_1: CSO_V attribute of A, A_2: VIET attribute of A, C_1: CSO_VIET attribute of C, C_2:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

EDE attribute of C, C_3: TULOAI attribute of C, CSE: Ede index, CSV: Vietnamese index, WV: Vietnamese word, WE: Ede word; EWE: Ede index, SP: part of speech, EX_{V-E}: Vietnamese-Ede bilingual example, note: noting for supplemented entries.

Method

+ Read(TG; WV, {WE}, SP, { EX_{V_E}})
+ SplitWord_E {WE} = (EWE_1, EWE_2,…, EWE_n)
+ SpiltExample_V_E {EX_{V-E}} = ({EX_{V-E1}}, {EX_{V-E2}}, …, {EX_{V-En}}) theo EWE_i
+ In turn EWV_i interacts with (A)
  - If WV ∈ A_1 then Read(A; CSV)
    - In turn EWE_i ∈ {WE} interacts with (A)

      If (CSV∉C_1) or (CSV∈C_1 & SP ∉ C_3) or (CSV∈C_1 & SP ∈ C_3 & EWE_i ∉ C_2) then Insert(C; CSV, EWE_i, SP, {EX_{V-Ei}})
  - If WV ∉ A_1
    - Insert(A; WV, note); Read(A; CSV)
    - Insert(C; CSV, EWE_i, SP, {EX_{V-Ei}})
+ Repeat Read(TG; WV) until EOF(TG)=true

b) The function block (4): interacting between Vietnamese corpus and Ede-Viet and Ede corpus

Input: the Vietnamese-Ede (C) corpus, the Ede corpus (B)

Ouput: the Ede-Vietnamese corpus (B)

Intermedium: A_1: CSO_V attribute of A, A_2: VIET attribute of A, C_1: CSO_VIET attribute of C, C_2: EDE attribute of C, C_3: TULOAI attribute of C, CSE: Ede index, CSV: Vietnamese index, WV: Vietnamese word, EWE: Ede index, SP: part of speech, EX_{V-E}: Vietnamese-Ede bilingual example

Method

+ Read(C; WE, SP, {EX_{V_E}}), Read(A; WV_{CSV})
+ WE Interacts (D)
  - If WE ∈ D_2 then Read(A; CSE)

    If (CSV∈B_1 & SP ∉ B_3) or (CSE∈B_1 & SP ∈ B_3 & WE ∉ B_2) then Insert(B; CSE, EW, SP, ConvertVE({EX_{V-Ei}},{EX_{E-Vi}})
  - if WE ∉ D_2
    - Insert(D; WE); Đọc(D; CSE)
    - Insert(B; CSE, EW, SP, ConvertVE({EX_{V-Ei}},{EX_{E-Vi}})
+ Repeat Read(C; VE) until EOF(C)=true

C. Results

Based on the Vietnamese-Ede interactive model and the Ede-Vietnamese dictionary, we have built the tool for the function blocks (1) and (2) to develop the Vietnamese-Ede and Ede-Vietnamese corpus with words and phrases and part of speech tagging. The entries quantity in the corpus shows in table II.

TABLE II. THE STATISTIC ENTRIES QUANTITY IN THE CORPUS

Corpus	no ambiguous words	ambiguous words
Ede-Vietnamese	9,306	10,744
Vietnamese-Ede	11,366	17,980

In the interactive process between the Ede-Vietnamese corpus and the Vietnamese corpus, the function block (2) has aligned the Vietnamese entry with the Ede entry and supplemented the Vietnamese entries into the Vietnamese corpus. The result of the function block (2) shows in table III.

The Ede entries have been checked by the Ede word segmentation tool on maximum matching method basing on the Ede-Vietnamese vocabulary corpus. The Ede language lessons and stories [13], [14] have been used the input of this tool. The result of checking has returned the entries that are not in the Ede-Vietnamese corpus. We checked and found that these entries are primarily the proper noun, correct misspellings and lacking.

TABLE III. THE RESULT OF THE BLOCK (2)

Vietnamese corpus	no ambiguous words	ambiguous words
Inheriting of the VLSP theme	31,242	34,038
The results of the fuunction block (2)	34,384	37,665
Alignment Vietnamese-Ede	11,366	17,980

IV. CONCLUSION

Criteria of developing the Vietnamese-Ede corpus includes unity using the Unicode fonts, building the infrastructure for Ede language processing and sharing in other research purposes. We proposed the conversion system raw data into structured data basing on the Vietnamese-Ede interactive model in building the Vietnamese-Ede vocabulary corpus. According to this system, we have built the functional blocks (1) and (2) for developing the Vietnamese-Ede vocabulary corpus. Our results have been the Vietnamese-Ede and Ede-Vietnamese vocabulary corpus that can be easily converted to other formats in order to share for other research purposes.

This system can be expanded to develop the bilingual vocabulary corpus for Vietnamese-other ethnic minorities.

This study has contributed to developing the infrastructure for Ede language processing in particular and the ethnic minority languages in general. In order to improve the quality of the corpus, we are going to continue designing the Vietnamese-Ede and Ede-Vietnam dictionary website based on the created corpus. Through this website, the quality of the corpus will be completed with the help and suggestions of the Ede language experts.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

REFERENCES

[1] Dak Lak Department of Education, *Vietnamese-Ede dictionary* (vol 1, 2), Vietnam Education publisher, 1993.

[2] Doan Cong Que, Scientific research theme on *Building Vietnamese-Ede dictionary*, http://www.baomoi.com/bao-cao-de-tai-khoa-hoc-xay-dung-tu-dien-dien-tu-viet-ede/c/1137624.epi.

[3] Doan Van Phuc, Ta Van Thong, *Ede language grammar*, Vietnam Education publisher, 2011.

[4] Hoang Thi My Le, Phan Huy Khanh, Souksan Vilavong, *Using Unicode in Encoding the Vietnamese Ethnic Minority languages, applying for the Ede language*, Proceedings of the Fifth International Conferece KSE 2013.

[5] Marek Rei, *Minimally supervised dependency-based methods for natural language processing*, https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-840.pdf

[6] Nguyen Thi Tuyet, *Building English-Vienamese-Ede dictionary*, C*omputer science master thesis*, Da Nang University, 2004.

[7] Phan Hong, Nguyen Van Thu Nay H'Ban, Y Dlông Nie, Dieu Linh, Nguyet Minh, *Ede-Vietnamese bilingual story* (vol 1, 2, 3), Vietnam Education publisher, 2013.

[8] Phan Thi Thu Nhan, *Building the Ede-Vietnamesecorpus in the Ede language processing*, C*omputer science master thesis*, Da Nang University, 20013.

[9] Ta Van Thong, "*Ede-Vietnamese dictionary*" theme, 2014.

[10] Vu Xuan Luong, *Building corpus applies for analysing, natural language processing, compiling dictionary*

[11] Y-Ha Nie Kdam, H'Mi Čil, *Klei Êđê* (Vol 1, 2, 3), Vietnam Education publisher, 2007.

[12] *VLSP theme*, http://vlsp.vietlp.org:8080/demo/?page=home.

[13] *Ede-Vietnamese dictionary*, http://giaoan.violet.vn/present/show/entry_id/9339030

[14] *Unicode*, http://vi.wikipedia.org/wiki/Unicode.