Building a Vietnamese-Ede Machine Translation Based on the Bilingual Corpus

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

Phan Huy Khanh
Danang University of Technology
The University of Danang
Danang, Vietnam

***Abstract*—Solving the serious lack of information in ethnic minority languages, the paper proposes building a machine translation system from Vietnamese into ethnic minority language in a restrict context such as bulletins about weather forcast, forest fire warnings, policies and laws of the State, cultivation experience, husbandry experience, natural disaster prevention methods etc. Translation method of system bases on Vietnamese-ethnic minority corpus in which the entries are indicated a domain for ethnic minority language terminology. This method has also contributed to solving the ambiguity of the word boundary, homonym word and polysemous word. Applying this method, a machine translation system from Vietnamese into Ede language has been built for translating the weather bulletins. A Vietnamese-Ede bilingual corpus with Ede language terminology in natural, geographical, hydrometeorology, weather forecast is used in translation system. The corpus has open structure and the major axis is in Vietnamese. The multilingual corpus is easily built by adding other ethnic minority languages such as: Cham, Ka Tu, Jarai, Muong etc.**

Keywords—machine translation; enthnic minority; bilingual corpus; ambiguity; Ede language

I. INTRODUCTION

According to the 1999 population and housing census results of the Statistics Documentation Center, General Statistics Office of Vietnam, ethnic Ede with the population of over 331 thousand, ranks No. 11 and accounts for 0.36% the country's population. In central Vietnam, Ede people live mainly in the provinces Dak Lak, Gia Lai, Khanh Hoa and Phu Yen. In some other countries, there are a few Ede people living in Cambodia, the United States, Canada and the Nordic countries [1][10].

Ede language belongs to Malayo-Polynesien linguistic family (Nam Dao). It has relations with the languages of the mainland Nam Dao [11][12]. The Ede Kpa language is the main dialect that being used in Tay Nguyen. On 12/02/1935, Governor-general of Indochina signed and recognized the writing system with Latin characters to use commonly for the EM in Tay Nguyen [8]. This alphabet is corrected many times and is called Ede alphabet because Ede is one of the ethnic minorities having the largest population in Tay Nguyen. However, there are not any website in ethnic minority (EM) languages. Even the website of the Committee for the EM Vietnamese CEMA [1], the websites of the locals where the ethnic people live are only in Vietnamese-Kinh language, or accompanied by English. The serious lack of information in aboriginal languages has made the economic and social of EM areas in Vietnam still underdeveloped and backward. Therefore, the problem of writing and cultural preservation and disseminating information in Ede language on the internet and media etc. for ethnic minorities and especially for Ede EM group is very urgent.

Currently, the Radio Voice of Vietnam and most of the local Radio and Television in the country have broadcasted in the EM languages. For example, the Radio and Television in Dak Lak broadcast programs in Ede language. The broadcast contents only help ethnic minorities have better understanding, more experience in economic development, animal husbandry, horticulture, forest, minerals, health care, preservation and promotion of their cultural values, maintaining border security, grasping the situations of climate, weather and soil. By means of the local Radio and Television, however, the staff training on understanding the culture and EM languages in general and Ede language in particular has been facing many difficulties. A procedure producing an Ede language broadcast is still manual and spends a lot of time and energy. Therefore, the support of the Information Technology in translating automatically the text from Vietnamese into Ede language is very necessary.

Through our survey, we found that a process producing the Ede language broadcast is still manual and spend a lot of time and energy. Almost, the broadcast content in Ede language is edited and translated from the newsletters, articles in Vietnamese language, reliability and authenticity between the root and transmits newsletter is not high. The production process Ede language broadcasts of the Radio Voice of Vietnam consists of 7 steps:

- Gathering the news, economic and social situation reports, production experiences, the categories of culture, music, health, weather forecast, the way to be rich
- Editing the contents in Vietnamese language by the editorial staffs
- Approving and signing by the leader of the Radio and Television
- Compiling manually the contents into Ede language by the staffs who are Ede people or can speak Ede language.
- Reading the contents in Ede language.
- Staging program
- Broading program

To improve the timeliness and efficiency of the Ede language broadcasting service of social and economic life for ethnic minorities, especially in weather forecasting activities, renewing step 4 of the above production process is necessary. That means translating the weather bulletins from Vietnamese into the Ede language using the advances of science and technology, applied information technology. Today, there are many the weather forecast websites, but there is not any website in EM languages for servicing the people in remote areas.

In Vietnamese language processing, machine translation problem has always been extremely difficult. However, we can build a machine translation system from Vietnamese into Ede language in a restrict context. That is the source text belonging to a narrow field, such as hydrometeorology (the weather forecast, forest fire warning, etc), crops, livestock, etc. These documents have a consistent structure, scientific content, language specific, less ambiguous. A sentence is almost identical, repeated. The machine translation system bases on the Vietnamese-Ede bilingual corpus with the entries that have been indicated the field of weather forecast for Ede language terminology.

This paper presents the solutions for building machine translation systems Vietnamese-Ede, which allow automatic translation of the weather bulletins from Vietnamese into Ede language in the Ede language broadcasting, Dak Lak Radio and Television.

II. ANALYZING THE WEATHER FORECASTS REPORTS IN EDE LANGUAGE

A. Characteristics and scripts of Ede language

In terms of type, Ede language is classified as isolating as Vietnamese. Unlike the Vietnamese-Kinh, is a monosyllabic language, Ede language is polysyllabic. For example, mơčrang (soi đèn) (to light up), lơkŭt (vắng mặt) (to absent), etc.

In terms of morphology, word base of Ede language is just morpheme, as well as capable of word forming independently. A rare characteristic in the languages in Vietnam is Ede language having many affixes that play a prefix role, such as *m*, *k* or infix, such as *n*, *d*. For example:

boh (quả trứng) (egg) → *mboh* (đẻ trứng) (to lay an egg)

truă (đậy) (to cover) → *ktruaw* (cái nắp) (lid)

kuôl (buộc) (to tie) → *knuôt* (nút áo) (button)

hrĭng (xâu lại) (to string) → *hdrĭng* (chuỗi) (string)

Ede scripts are mainly transcribed from common language (National language), so almost using vowels and consonant root is not changed [2]. In addition, Ede language also have more special vowels and compound consonants, that there is no in the Vietnamese language. For example [2]:

Special vowels: *ab*, *ad*, *aw*, *ar*, *al*, *aj*, *ah*, *êk*, *êth*, *êƀ*, *êb*, *êd*, *êđ, êdj,êj, ês, êr, êl, êñ, êg, êh, êy.*

Compound two consonants: *mb*, *mƀ, mm, mw, mt, mđ, md, mn, ms, mr, ml, mč, mj, mñ, my, mk, mg, br, bl, bh, ƀr, ƀl ƀh, pr, pl, tl, đr, đh, dr, dh, dl, ñh, jh, kp, kƀ, kb, km, kw, kt, kđ, kd,* kn, *ks*, *kr, kl, gr, hb, hm, hn, hđ, hd, hr, hl, hj, hñ, hy, hw, hg.*

Compound three consonants: *mpr*, *mpl*, *mƀr*, *mƀh*, *mbr*, *mbl, mbh, mtr, mtl, mđr, mdr, mdl, mđh, mdh, mnh, mjh, mñh,* *mkr*, *mkl*, *mgr*, *mgh*, *mhr*, *mhl*, *kpr*, *kpl*, *kƀr*, *kbr*, *kƀl*, *kbl*, *kƀh*, *kmr*, *kml*, *kmh, ktr, ktl, kđr, kdh, kdl, kdr, knh, kmh, klh, kñh,* *hml*, *hdr*.

The Ede sentences built from single words, compound words, derivative and repeat words. Compound words are formed from grafting the morphemes, which are capable of functioning as independent words. For example: Word *yang hruê* (mặt trời- sun) derives from grafting two single words: *yang* (thần, thánh- god) and *hruê* (ngày- day). Repeat word is a word in which syllable related phonetically together. For example: *răng* (rối- tangle), *êmit êmang* (yên tĩnh- quiet), *siam siăn* (đẹp- beautiful).

B. Grammatical features of Ede language

Vietnamese and Ede languages have many similarities about using grammar rules, not having morphological metamorphosis. The syntax order of a Ede sentence is similar to Vietnamese sentence, especially in narrative sentence [2]. There are two types of simple sentence in Ede language, including:

Type 1: The subject (noun/verb) + Adjective / verb. For example: *Adiê hjan* (Trời mưa) (It rains)

Type 2: The subject (noun/verb) + Adjective/verb + The object. For example: *Amĭ tăp mdiê.* (Mẹ giã gạo.) (mother pounds rice.)

Complex sentences in Ede include many simple sentences, each simple sentence consists of words (single words or compound words) paired together in the same order in Vietnamese. For example:

Mla| lu| knam|, leh| hjan| rưng khưng khah| dua tlâo| Anôk.

Đêm| nhiều| mây|, có| mưa| rải rác| vài| nơi.

(The night is cloudy, and rain scattered few places.)

Mbruê| kâo| lei| êdeh wai jơ̆ng| kơ| anak.

Hôm qua| tôi| mua| xe đạp| cho| con.

(I bought bicycle for child yesterday.)

The Ede language also has some characteristics that are different from the Vietnamese. Adverbs can stand in front the adjectives, for example: *thâo snăk* (*thâo*: giỏi- good) (*snăk:* rất- very). In the question of Ede language, question word always put at the beginning of a sentence. For example: *Ti anôk sang ih?* (nhà anh ở đâu?- where is your house). *Ti anôk* (ở đâu- where).

C. Analysing the handmade translation

Each weather forecast report usually has two contents: daily weather and hydrological risk weather forecast such as: wind storms and tropical low pressure, cold, heat, floods, tornadoes, rain rocks, earthquakes, tsunamis etc. After collecting and sorting the weather bulletin in Vietnamese language, the translators translate into Ede language, before reading, compiling and arranging programs.

The analysis results of the manual translation show that the order of words in the sentence of the Ede language almost similar to the one of Vietnamese but there are some distinctive characteristics leading to the phenomenon of ambiguity while translating. To ensure the Ede people in different regions can hear, understand, accept the translation information in many areas, including weather, it is necessary to handle the following cases:

Differences between single words and compound words: From one word in Vietnamese language, we can translate into a word or a phrase in Ede language. From single word in Vietnamese language, we can translate into a compound one in Ede language, or vice versa. For example, the pairs of word Vietnamese/ Ede as following:

Thời tiết */ Adiê* (weather)
U ám */ gâm* (overcast)
Nhiệt độ */ Hnơ̆ng hlơr ê-ăt* (temperature)
Gió nồm / *angĭn mơ̆ng yŭ* (south *wind)*
Bão / *angĭn êbŭ* (storm)
Đảo / *plao êa* (island)
Và / *leh anăn* (and)

The order of words in a sentence: processing the case of *the* adverb standing behind the adjective. For example, “Trời mưa rất to.- It rains very heavy” translation into *Adiê hjan ktang snăk* (rất- very).

III. BUILDING A VIETNAMESE-EDE MACHINE TRANSLATION SYSTEM

We choose the machine translation method basing on the Vietnamese-Ede bilingual corpus. It contains the entries that are indicated the field of weather forecast with Ede terminology. The process of machine translation is word matching, do not require analyzing syntax and semantics. After analyzing to split into independent sentences and to continue to split sentence into words. Each word is matched to the entries in corpus to retrieve translation results by replacing 1-1. The accuracy of the method depends on the Ede terminology of the entries are stored in corpus. Figure 1 shows the model architecture of the system.

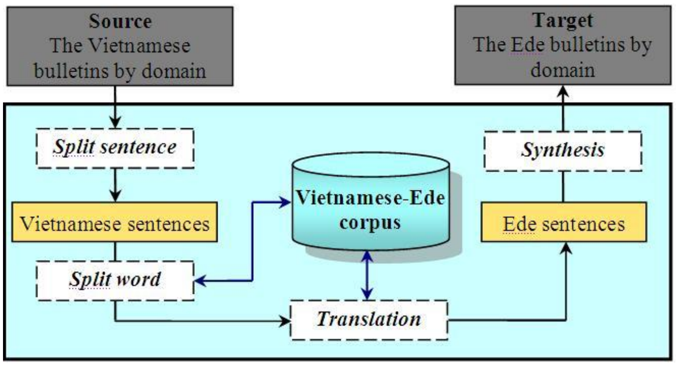


Fig. 1. The model architecture of the system

The machine translation process of weather forecast reports from Vietnamese into Ede language includes the following steps:

- Building Vietnamese-Ede bilingual corpus, mainly in the field of hydrometeorology, weather forecast
- From the Vietnamese weather forecasts reports, analyzing to split into independent sentences and to split words in sentence basing on the Vietnamese entries in Vietnamese-Ede bilingual corpus
- From the split words in Vietnamese, the machine translation system searches the words in Ede language from the Vietnamese-Ede bilingual corpus, processes the translation situations of the order of words, proper nouns, numbers, symbols etc.
- Synthesizing the result sentences to received the text of weather forecast in Ede language.
- Checking the machine translation results and interacting with the user to receive the results.

A. Criteria for the Vietnamese-Ede bilingual corpus

A Vietnamese-Ede bilingual corpus is built according to the following criteria:

- The words in a corpus have meanings and are represented in the scientific documents.
- The field of the corpus relates to hydrometeorology, weather forecast.
- The documents are selected for the building the corpus, which relate to the weather bulletins.
- The Vietnamese-Ede bilingual corpus is done alignment according to level of Vietnamese word with Ede language word which is indicated the field of hydrometeorology, weather, natural, geographical.

The Vietnamese-Ede bilingual corpus is saved in the computer with Unicode fonts (supports Vietnamese). This is the problem that the previous EM corpus has not been mentioned.

B. Developing a Vietnamese-Ede bilingual corpus

Basing on the Vietnamese monolingual corpus with segmented word [9], we propose a method indicating the field for the Ede terminology according to the above criteria.

The field indicating method for Ede terminology:

Input

The Vietnamese monolingual corpus with 31,248 words
Ede-Vietnamese dictionary (text file with TNKey font) [5]
Vietnamese-Ede dictionary (book)
The Vietnamese document files by field

Output

The Vietnamese-Ede bilingual corpus is indicated the field for Ede terminology.

Method

- Building the macro of visual basic for application in Microsoft Word [6] to format the Vietnamese monolingual corpus into the corpus with the table structure

- Using CEDU program [7] to convert the text file of the Ede-Vietnamese dictionary with TNKey font into Unicode font

- Building the macro of visual basic for application in Microsoft Word to format the text file of the Ede-Vietnamese dictionary into the corpus with the table structure

- Building the field indicating method for Ede terminology program (DIMET–Domain Indicating Method for Ede Terminology) by interacting in the Vietnamese monolingual corpus to create a Vietnamese-Ede bilingual corpus with the field indicating for the Ede terminology

The operation of the DIMET tool:

DIMET tool segments word from the Vietnamese document files by domain. The word segmentation method is applied in DIMET that is the longest matching method because DIMET is inherited the Vietnamese corpus with segmented words.

DIMET interacts with the Vietnamese monolingual corpus to indicate the field for the segmented word with the context of the document. The indicating field helps user selecting the Ede terminology in the alignment operation.

The words do not belong to the Vietnamese corpus. They will also be save to the other corpus. After that, the user checks and saves them into the Vietnamese corpus. This contributes to improve the quality of the corpus.

With the function of the Ede word alignment, the user can select the Ede word in the Ede-Vietnamese corpus according to the indicated domain of the Vietnamese word or input by hand for the Ede word not in the Ede-Vietnamese corpus. This operation contributes also to solve the ambiguity of the synonym word but not of homonym. For example, the "*vùng*" in Vietnamese corresponds to four words in Ede language: *Alŭ, Wăl, Ễñah and Kluh.*. According to hydrometeorology and weather forecast, the user selects "*Alŭ*".

For the Vietnamese words which do not belong to the Ede-Vietnamese corpus will be updated manually based on a Vietnamese-Ede dictionary [3].

To solve partly the ambiguity, have chosen the word segmentation for the documents in the restrict context. These documents belong to the specialized domain with the simple sentences, which are less ambiguous and not abstract. For example, the documents about the forecast weather, the hydrometeorology, the cultivation techniques, the animal husbandry methods, forest fire warning, etc.

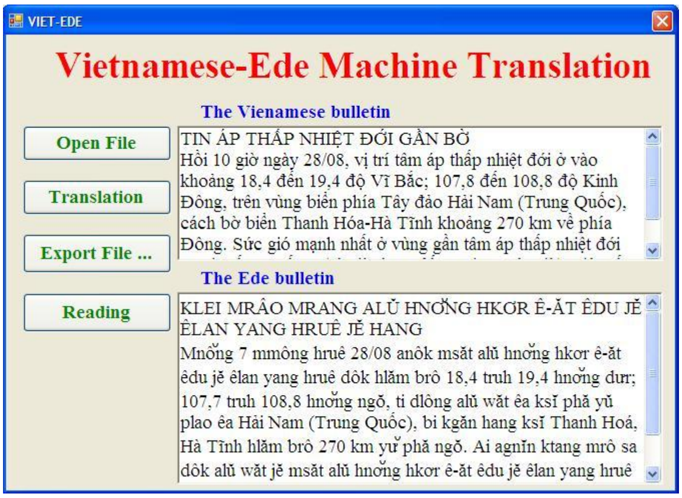
By DIMET, we have tested with 800 files about hydrometeorology, weather forecast and achieved the result: creating a Vietnamese-Ede bilingual corpus with 31,248 Vietnamese entries and 2,500 Ede entries that are indicated terminology of natural, geographical, hydrometeorology, weather forecast and many other popular words.

C. *Testing*

With these initial results, our Vietnamese-Ede bilingual corpus contains 2,500 Ede entries of hydrometeorology, weather, natural, geographical and many other popular words. The bulletins are treated as pure text, no pictures and diagrams.

The interface of the Vietnamese-Ede machine translation system is showed in the figure 2.

Before using the machine translation results, the manual translation often depends on qualifications and experience ... of translators. Therefore, it was often wrong and had many spelling mistakes, inconsistencies between the translation results. Meanwhile, the machine translation results are very quick, accurate and without errors or spelling mistakes, and always consistent.



TIN ÁP THẤP NHIỆT ĐỚI GẦN BỜ
Hồi 10 giờ ngày 28/08, vị trí tâm áp thấp nhiệt đới ở vào khoảng 18,4 đến 19,4 độ Vĩ Bắc; 107,8 đến 108,8 độ Kinh Đông, trên vùng biển phía Tây đảo Hải Nam (Trung Quốc), cách bờ biển Thanh Hóa-Hà Tĩnh khoảng 270 km về phía Đông. Sức gió mạnh nhất ở vùng gần tâm áp thấp nhiệt đới

KLEI MRÂO MRANG ALŬ HNỞNG HKƠR Ê-ĂT ÊDU JĔ ÊLAN YANG HRUÊ JĔ HANG
Mnống 7 mmông hruê 28/08 anôk msăt alŭ hnởng hkơr ê-ăt êdu jĕ êlan yang hruê dôk hlăm brô 18,4 truh 19,4 hnởng dưr; 107,7 truh 108,8 hnởng ngŏ, ti dlông alŭ wăt êa ksĭ phă yŭ plao êa Hải Nam (Trung Quốc), bi kgăn hang ksĭ Thanh Hoá, Hà Tĩnh hlăm brô 270 km yư̆ phă ngŏ. Ai agnĭn ktang mrô sa dôk alŭ wăt jĕ msăt alŭ hnởng hkơr ê-ăt êdu jĕ êlan yang hruê

Fig. 2. Vietnamese-Ede machine translation system

The following example is a result of our machine translation.

Văn bản nguồn: TIN ÁP THẤP NHIỆT ĐỚI GẦN BỜ

Hồi 10 giờ ngày 28/08, vị trí tâm áp thấp nhiệt đới ở vào khoảng 18,4 đến 19,4 độ Vĩ Bắc; 107,8 đến 108,8 độ Kinh Đông, trên vùng biển phía Tây đảo Hải Nam (Trung Quốc), cách bờ biển Thanh Hóa-Hà Tĩnh khoảng 270 km về phía Đông. Sức gió mạnh nhất ở vùng gần tâm áp thấp nhiệt đới mạnh cấp 6, cấp 7 (tức là từ 39 đến 61 km một giờ), giật cấp 8, cấp 9.

Source Documents: INFORMATION TROPICAL DEPRESSION NEAR THE SHORE

In 10 hours on 06/10, the position in mind tropical depression at bout 18.4 to 19.4 degrees north; 107.8 to 108.8 East longitude, on the west coast of Hainan Island (China), off the coast of Thanh Hoa - Ha Tinh about 270 km to the east. Highest wind speed near to the center and tropical low pressure energy level 6, level 7 (ie from 39 to 61 km per hour), the shock level 8, level 9.

Manual translation: KLEI MRÂO MRANG ALŬ HNỞNG HKƠR Ê-ĂT ÊDU JĔ ÊLAN YANG HRUÊ JĔ **HANG KSĬ**

Mnống 7 mmông hruê 28/08 anôk msăt alŭ hnởng hkơr ê-ăt êdu jĕ êlan yang hruê dôk hlăm brô 18,4 truh 19,4 hnởng dưr; 107,7 truh 108,8 hnởng ngŏ, ti dlông alŭ wăt êa ksĭ phă yŭ plao êa Hải Nam (Trung Quốc), bi kgăn hang ksĭ Thanh Hoá, Hà Tĩnh hlăm brô 270 km yư̆ phă ngŏ. Ai agnĭn ktang mrô sa dôk alŭ wăt jĕ msăt alŭ hnởng hkơr ê-ăt êdu jĕ êlan yang hruê ktang gul 6, gul 7 (jing dởng mởng 39 truh 61 km sa mmông), kplăk gul 8, gul 9.

Results of machine translation: KLEI MRÂO MRANG ALŬ HNỞNG HKƠR Ê-ĂT ÊDU JĔ ÊLAN YANG HRUÊ JĔ **HANG**

Mnống 7 mmông hruê 28/08 anôk msăt alŭ hnởng hkơr ê-ăt êdu jĕ êlan yang hruê dôk hlăm brô 18,4 truh 19,4 hnởng dưr; 107,7 truh 108,8 hnởng ngŏ, ti dlông alŭ wăt êa ksĭ phă yŭ plao êa Hải Nam (Trung Quốc), bi kgăn hang ksĭ Thanh Hoá, Hà Tĩnh hlăm brô 270 km yư̆ phă ngŏ. Ai agnĭn ktang mrô sa dôk alŭ wăt jĕ msăt alŭ hnởng hkơr ê-ăt êdu jĕ êlan yang hruê ktang gul 6, gul 7 (jing dởng mởng 39 truh 61 km sa mmông), kplăk gul 8, gul 9.

Comparing the results of the manual translations and machine translations, we realized that errors are mainly the spelling mistakes of manual translation, adding word by translator, for example "near shore" translated "near coast". Besides, the accuracy of the Ede terminology in a the Vietnamese-Ede corpus also contribute to improving the accuracy of the translation system.

IV. CONCLUSION

Solving the serious lack of information in the EM languages in general and the Ede language in particular we have built a Vietnamese-Ede machine translation system for the weather bulletins. Translation method of the system is based on a Vietnamese-Ede bilingual corpus with 31,248 Vietnamese entries and 2,500 Ede entries that are indicated terminology of natural, geographical, weather forecast, the hydrometeorology. The system was tested at Radio and Television of Dak Lak province. The initial results are very encouraging. Next, we will continue to improve the system, add more data, evaluate results of operations. This system is applied to build the website in Ede language containing the weather forecast categories.

Moreover, the results of the machine translation system is expanded into other fields such as forest fire warning, policies and laws of the State, farming experience, animal husbandry experience, natural disaster prevention, etc. The corpus has open structure and the major axis is in Vietnamese. The multilingual corpus is easily built by adding other ethnic minority languages such as: Cham, Ka Tu, Jarai, Muong etc.

REFERENCES

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Committee for Ethnic Minorities Vietnamese CEMA CEMA, http://cema.gov.vn/modules.php?name=Content&op=details&mid=498

[2] Dak Lak Department of Education, *EDE language Grammar*, the education Publisher (2011).

[3] Dak Lak Department of Education, Vietnamese-Ede dictionary (vol 2), the education Publisher (1993).

[4] Dinh Dien, Hoang Kiem. *Building an Annotated Parallel Corpus of English-Vietnamses*. Proceedings of International Conference on Natural Language Processing, ICON’04, India (2004).

[5] Ede-Vietnamese dictionary, http://giaoan.violet.vn/present/show?entry_id=9339030

[6] Khanh Phan Huy, *Using VBA macro programming tool to build text processing utilities*, The Journal of Science and Technology Danang University, No.10, (2005).

[7] Le Hoang Thi My, Khanh Phan Huy, *Using Unicode in Encoding the Vietnamese Ethnic Minority languages, Applying for the Ede Language,* In Proceeding of The Fouth International Conference on Knowledge and System Engineering, Springer, KSE2013, HaNoi (2013).

[8] Le Khac Cuong, *Researching ethnic minority language in VietNam*, In Proceeding of The Second International Conference on Researching and comparing the humanity Taiwan-VietNam, Taiwan (2013).

[9] Lưu Tuấn Anh, *Vietnamese Natural Language Processing*, http://viet.jnlp.org/

[10] Vietnam General Statistics Office, http://www.gso.gov.vn/default.aspx?tabid=407&idmid=4&ItemID=1346

[11] Y Čang Niê Siêng (1979), Ede - *Rade Vocabulary*, The Vietnamese minority languages bookshelves of The Summer Language Institute , XIV

[12] Y Čang Niê Siêng, Y ČôC Mlô, *Hdruôm Hră Hriăm Êđê*, Dak Lak Department of Education (2007).