

# Developing Database of Vietnamese Abbreviations and Some Applications

Nguyen Nho Tuy<sup>1(✉)</sup> and Phan Huy Khanh<sup>2</sup>

<sup>1</sup> VNPT Da Nang, Danang, Vietnam  
nhotuy68@gmail.com

<sup>2</sup> University of Science and Technology – The University of Danang, Danang, Vietnam  
khanhph29@gmail.com

**Abstract.** Abbreviations (CVT) in documents are widely used in various fields and in many languages including Vietnamese. In fact, currently, abbreviations are often repeatedly and unclearly used, demand for abbreviation use is increasing that requests a database of plentiful abbreviations which is saved and used conveniently, is easy to update and consistently exploited. In this article, we propose an opening solution in order to develop a database of Vietnamese abbreviations for many purposes of use during processing language and exploiting database.

**Keywords:** Abbreviation · Acronym · Database · Index abbreviation

## 1 Introduction

Abbreviations have been often used in daily life and widely used in almost all of languages in the world so far, including Vietnamese. In newspapers, magazines, we often see common abbreviations such as TŨ (Trung ương), UBND (Ủy ban nhân dân), and also English abbreviations such as WTO (World Trade Organization, etc. Owing to abbreviations, all texts are shorter and simpler but express more capacity of information. The fact that abbreviations are often used makes the abbreviation system increasingly diversified and abundant. On the one hand, users (NSD) have many abbreviations to choose and use, on the other hand the users also run into a lot of difficulty in finding, searching its meanings and proper using of such abbreviations.

With regard to abbreviations, there are some dictionaries today such as “Từ điển giải nghĩa thuật ngữ Viễn thông” (Dictionary of telecommunication), “Thuật ngữ viết tắt Viễn thông” [8] (Dictionary of abbreviations in telecommunication; websites of CVT [16], but mainly in foreign languages. The need of abbreviations is higher and higher, wider and wider and indispensable, especially brands, trademarks, etc.

Contents of this article include: Firstly, we present information about abbreviations, history of abbreviation development, principle to create abbreviations, classification of abbreviations and influential factors in abbreviation generation. Next, we present a solution to developing database of abbreviations, statistically assess results and give solution of abbreviations. The final is conclusions.

## 2 Information About Abbreviations

### 2.1 Definition and Terms

The term “chữ viết tắt” (In English: abbreviation) has not been appeared in Common Vietnamese dictionary in current market<sup>1</sup> including in “Từ điển Bách khoa Việt Nam” Vol. 1 (Letters A-Đ<sup>2</sup>), but is very familiar in daily life.

We often see abbreviations or acronyms. They are used for generating abbreviations that are different from common written languages; abbreviations are used when we have to repeatedly write a word, phrase, sentence or paragraph for convenience [8]. For a long time ago, people used abbreviations to inscribe on stone, wood, etc. in order to save time, force and material. According to Manuel Zahariev [14], abbreviations are originated in Ancient Greek, *acronym* includes *akron* (the last or first one) and *onoma* (name or voice). According to some English dictionaries, abbreviations are the way to form new shorter words by using initial letters, or last letters or any letters of a word. For example, UNESCO stands for “United Nations Educational, Scientific and Cultural Organization”, etc.

We also see abbreviations in short form, it means that a phrase or a paragraph is abridged in some characters, is extracted, chose or replaced any part to form a set of new characters, in order to writing and saying are more convenient. For example, abbreviations are used for geographical areas, for example, Thanh Land, Nghe Land, Quang Land, etc.

In the progress of Internet explosion, general written languages have been developed towards a new direction owing to the use of various abbreviations and conventional signs. For example, in English, email, messages, IMHO stands for “in my humble opinion”, comic signs ☺, ☹, U (you)... The use of abbreviations in fields of information technology and communication today on one hand makes users beneficial, on the other hand, such diversity of abbreviations also troubles the users.

### 2.2 History of Abbreviations

Abbreviations have been widely used for a long time ago in foreign countries. For example, SPQR stands for “Senatus Populusque Romæ” and have been appeared for nearly 2000 years [14], QED stands for “Quod Erat Demonstrandum” in “Ethica More Geometrico Demonstrata” of a philosopher, Benedictus de Spinoza (1632–1677).

In Vietnam, today, there are some researches into Vietnamese abbreviations [4, 12], but such researches are not complete and systematic, although Vietnamese abbreviations have been early formed. The formation of “chữ Nôm” (an ancient ideographic vernacular script of the Vietnamese language) since the 18<sup>th</sup> century has been other way to write “chữ Hán” (Chinese writing), replacing “chữ Hán” after nearly one thousand years of occupation and colonization by the Han [2, 3]. In the “chữ Nôm”, each “chữ Nôm” is

<sup>1</sup> Vietnamese-English dictionary, Bui Phung, published by global publishing house in 1998.

<sup>2</sup> Vietnam encyclopedia compilation steering council compiled. Vietnam encyclopedia compilation center published in 1995.

square, is formed by putting “chữ Hán” together in the form of onomatopoeia, pictographic or reducing characters, abbreviation. For example, the Chinese writing 共 (total) is reduced its characters into “chữ Nôm” 𠂇 (khặng), “chữ Hán” 𠂇 is reduced its characters into “chữ Nôm” 𠂇 (lâm).

When Vietnamese national language (Current Vietnamese language) had been widely used, abbreviations have been used. The pen name C.D. standing for Chương Dân is official name of Phan Khôi in “Đông Pháp Thời Báo” in 1928. Today, Vietnamese abbreviations are being used increasingly widely in many fields.

Many authors think that Vietnamese abbreviations refer to a grammar [1, 9, 10]. According to Prof. Nguyen Tai Can, we “*use abbreviation in form of one syllable rather than in form of initial letters. The acronyms such as DT (danh từ), VN (Việt Nam), HTX (hợp tác xã), etc. only are used in writing documents*”. Although there are many views of the use of abbreviations, abbreviations are existing as an internal part of Vietnamese language, and there are many abbreviation applications in communication, text processing, data exploitation [5], etc.

### 2.3 Principles of Abbreviation Generation

Based on the results of analysis, the demand and current status of the abbreviation use in daily life, we proposed 07 Principles of abbreviation generation as detailed [4], and now, we supplement 2 new Principles of abbreviation generation (Principles 8 and 9).

1. 7 Principles of abbreviation generation that have been developed: Principle of abbreviation; principle of word connection; principle of short connection by meaningful words; principle of sub-letters; principle of connection of foreign languages; principle of borrowing of abbreviations in foreign languages; principle of random abbreviation.
2. 2 new principles include:
  - (1) Principle of encrypted abbreviations:
 

In many fields and sections, reminiscent abbreviations are used in conformity with a predefined rule to encrypt the phrase. All encrypted abbreviations often must satisfy:

    - i. Encrypted abbreviations often are issued by an organization with scope of use and application.
    - ii. Encrypted abbreviations are unique and unduplicated to avoid ambiguity.
    - iii. Encrypted abbreviations are often used new characters according to a predefined rule.

For example, lists and tables in database, list of national codes, regional codes, sectional codes, and codes of telecom fiber optic cables, etc.
  - (2) Principle of abbreviations in database:
 

According to studying in theories of searching problems, relevant practical results and the efficient use of abbreviations, we propose some principles applying index abbreviations in order to search data in large database:

    - i. Abbreviations only are used English letters (not Vietnamese words) and digits 0...9

- ii. Don't use special characters: punctuation marks, space (SP)
- iii. Abbreviations are reminiscent, short, not unduplicated, and not unclear: Users immediately image abbreviations after determining request for information searching.
- iv. Implement index of database on the established fields of abbreviations.

The Fig. 4 shows the results of developing database by applying abbreviations for index (called *index abbreviations*) and phone subscriber lookup in the Switchboard 108 VNPT Da Nang [5].

## 2.4 Influential Factors in the New Abbreviation Generation

According to the field survey, we propose 4 influential factors in the generation of new abbreviations, particularly:

**Number of characters:** Abbreviations shall not be too long. In general, common number of characters of an abbreviation should not more than 18 characters.

**Marks in Vietnamese language:** Avoid vowel with mark such as  $\hat{a}$ ,  $\check{a}$ ,  $\sigma$ ,  $\hat{e}$ ..., don't use grave, acute, question mark and dot below in abbreviations in order to avoid misunderstanding, difficult to speak.

**Spiritual factors for East Asians:** Select number of characters of an abbreviation. Avoid number 2, number 4 or avoid number of characters of an abbreviation according to the conception of *birth*, *old age*, *illness*, *death*. In order to generate the word "birth", the number of characters of an abbreviation shall be 5, 9, 13, etc., and in to generate the word "old age", the number of characters of an abbreviation shall be 2, 6, 14, etc.

**Syllable:** Select abbreviations so that when being read, such abbreviations are formed opening and deep hollow notes. People often choose *a*, *ô*, *i*, or *ex*, *ec*, rather than  $\hat{e}$ ,  $\sigma$ .

Two last factors often are specially considered when finding abbreviated name of enterprises, companies, brands, trademarks, organizations, projects, etc.

## 2.5 The Use of Abbreviations

Generally, users shall define or explain all abbreviations in documents. There are two cases as below:

**Using available abbreviations:** Abbreviations are defined and explained previously, or commonly used, not unclear.

**Using new abbreviations:** Defining and using abbreviations right after initial appearance in documents in the form of:

<Complete phrase > (< Abbreviation >)

The above principles of abbreviation generation allow us to refer 05 signs of abbreviations in a Vietnamese document, particularly:

- (1) Abbreviations are placed in brackets (...), or placed after the phrases: “viết tắt là”, “viết tắt”, “gọi tắt là”... (hereinafter referred to as..., hereinafter called, etc.) when the abbreviations are defined initially.
- (2) Abbreviations are capital letters (lowercase in normal letters)
- (3) Abbreviations include special letters or marks: *and (&)*, *cross mark (/)*, *dash (-)*, *dot (.)*, *space*, and letters and digits, etc.
- (4) Abbreviations are words whose number of characters may be 18.
- (5) Vietnamese abbreviations shall not include vowels *â, ă, ơ, ê, ô...* don't use marks such as *grave, acute, question mark and dot below*.

## 2.6 Ambiguity of Abbreviations

Ambiguity of abbreviations is not rare, the ambiguity is formed by natures: difficult to understand abbreviations, arbitrary abbreviations, not complying with rules, difficult to define meaning of abbreviations:

For example: VH: Văn hóa, Văn học; Abbreviations are local, uncommon: Cao Xà Lá: Cao su, Xà phòng, Thuốc lá; Phối kết hợp: Phối hợp, kết hợp; not complying with rules: SKZ: súng không giật/z...

The principles of abbreviation generation 1–8 often cause Ambiguity. The principles 8, 9 do not cause Ambiguity within scope and application of abbreviations. However, Vietnamese abbreviations in general have the following characteristics:

- (1) Difficult to define meaning of abbreviations due to the way of writing
- (2) Abbreviations often are formed for easy to speak, to remember and convenient, thus abbreviations are often concise and polysemous.
- (3) Abbreviations continuously change; the formation of language @ and use of foreign language abbreviations make abbreviations increasingly plentiful and diversified;

## 2.7 Classification of Abbreviations

There are many methods in classifications of abbreviations, basing on field of use, site, or alphabet, etc. In article published in 2006 [4], we recognized 9 fields; and by now, with classifications of abbreviations up on field of use, we recognized the 12 main fields (Table 1).

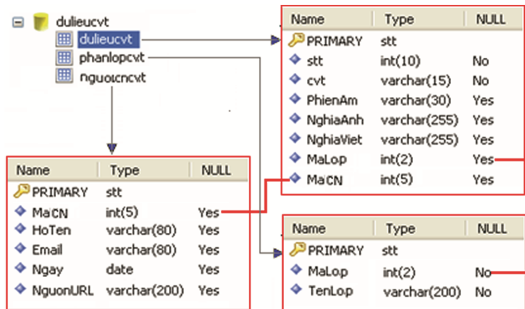
**Table 1.** Statistics of database of abbreviations.

Cate-gory	Fields of abbreviations	Manual update	Automatic update	Total	% of auto-update
1	Information technology and communication	754	350	1104	32 %
2	Government, political and social organizations	301	120	421	29 %
3	Science, technology and engineering	273	253	526	48 %
4	Military	202	120	322	37 %
5	Medicine	253	255	508	50 %
6	Education	301	2378	2679	89 %
7	Finance, trade	403	140	543	26 %
8	Environmental resources	163	130	293	44 %
9	Community communication	121	125	246	51 %
10	Religion	0	150	150	100 %
11	Proper name	0	75	75	100 %
12	Other	0	120	120	100 %
	<b>Total</b>	<b>2771</b>	<b>4216</b>	<b>6987</b>	<b>60 %</b>

### 3 Developing Database of Vietnamese Abbreviations and Some Applications

#### 3.1 Model of Database

We develop database (database) for abbreviations, including 3 tables of DULIEUCVT (data of abbreviation), PHANLOPCVT (classification of abbreviation) and NGUOICNCVT (editor of abbreviation) with relations as figure below (Fig. 1).



**Fig. 1.** Relations of database of abbreviations.

Table DULIEUCVT contains information including: order of abbreviations, field of abbreviations, phonetic field to easily read the fields, field of meanings (explanations) in English and fields in Vietnamese, fields of layer codes and fields of updated codes which are outer locks connecting to two databases accordingly. Table DULIEUCVT contains all possible abbreviations for exploit and continuous update. Table PHAN-LOPCVT enlists layers of abbreviations including code and name of layer.

### 3.2 Update of Abbreviations

We use different sources of abbreviations to update the database. The update process is conducted in the two main steps:

#### *Step 1: Manually update*

Directly update into WinWord documents from different sources like books, newspapers, magazine, legal documents, scientific reports or practicality, etc.

#### *Step 2: Auto-update from internet*

Base on the result of step 1, continue to automatically enrich database of abbreviations from Internet environment. Base on the aware signals of abbreviations in a document, we draw new abbreviations to supplement the database. We also develop a search engine for abbreviations as the introduced principles in [13]. Algorithm describes the operations of search engine for abbreviations [4] in the Internet environment is shown as below:

```

Algorithm :
Input : Address URL
Output: Data of abbreviations in table TUDONGCVT
Open intermediary database
Define operative URL
Save URL in intermediary database
Activate abbreviation counter
Repeat
  Open a file HTML
  Read content respectively HTML
  Dissect data (remove space and tags HTML)
  Find abbreviations basing on aware signals
  If found abbreviations Then
    Check whether abbreviations exist or not?
    If abbreviations exist CVT Then
      Increase abbreviation counter
    Else
      Save abbreviations and put the corresponding
      value by 1
      Extract sentences containing abbreviations
    End If
  End If
Until no more HTML

```

After collecting abbreviations from files HTML, continue to classify abbreviations to add in database.

**Step 3: Compile data for abbreviations**

This phase needs the involvement of experts to retrieve, refine and edit data. The updating process will include test and warnings for repeat of abbreviations or repeat in meanings.

Interface of Admin website for update and edit of abbreviations will be developed as the Fig. 2.

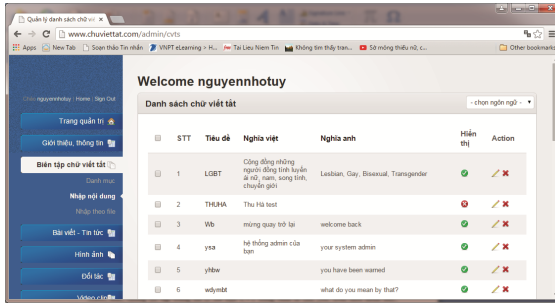


Fig. 2. Interface of Admin website for edit and update of database of abbreviations.

**3.3 Statistics of the Result**

Formerly, we focused on the update of abbreviations in English. By now, we have enlisted the number of existing English and Vietnamese abbreviations in database as follows:

According to the statistics result, much data of abbreviations is barely updated; abbreviations continuously change. Particularly, education field owns lots of abbreviations, mainly relating to code of colleges, professionals and specialties.

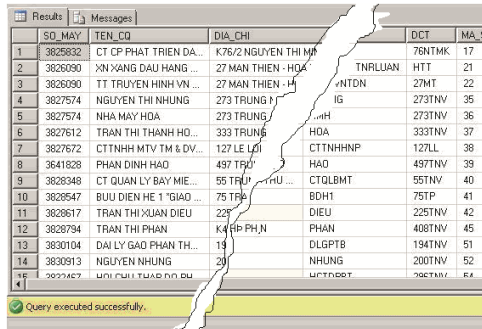


Fig. 3. Interface of website for abbreviation exploitation.



### 3.4 Website for Management and Use of Abbreviations

We build a website [www.chuviattat.com](http://www.chuviattat.com) (Fig. 3) containing database of abbreviations and managing online searching of abbreviations in Vietnamese and English to serve users intensively.



	SO_MAY	TEN_CQ	DIA_CHI	DCT	MA_SQ
1	3825832	CT CP PHAT TRIEN DA...	K76/2 NGUYEN THI MA...	76NTMK	17
2	3826090	XN XANG DAU HANG...	27 MAN THIEN - HOA...	TRNLLIAN HTT	21
3	3826090	TT TRUYEN HINH VN...	27 MAN THIEN - HA...	HTNTON 27MT	22
4	3827574	NGUYEN THI NHUNG	273 TRUNG M...	IG 2731NV	35
5	3827574	NHA MAY HOA	273 TRUNG M...	PH 2731NV	36
6	3827612	TRAN THI THANH HO...	333 TRUNG B...	HOA 3331NV	37
7	3827672	CTTNHH MTV TM & DV...	127 LE LOI	CTTNHHNP 127LL	38
8	3641828	PHAN DINH HAO	497 TRU...	HAO 4971NV	39
9	3828348	CT QUAN LY BAY MIE...	55 TRU...	CTQLBMT 551NV	40
10	3828547	BUU DIEN HE 1 'GIAO...	75 TRU...	BDH1 75TP	41
11	3828617	TRAN THI XUAN DIEU	225 TRU...	DIEU 2251NV	42
12	3828794	TRAN THI PHAN	K4 PH PH N	PHAN 4081NV	45
13	3830104	DAI LY GAO PHAN TH...	19 TRU...	DLGPTB 1941NV	51
14	3830913	NGUYEN NHUNG	20 TRU...	NHUNG 2001NV	52

Fig. 4. Result of building database for phone subscriber lookup through Switchboard 108 VNPT-Da Nang

### 3.5 Abbreviation Transfer in Database

Beforehand, we transfer abbreviations from Vietnamese into English for storage which helps the online searching of foreigners, users, comparison, and definitions and avoids the repeat of abbreviations. Subsequently, the fully edition will be published in book.

### 3.6 Abbreviation Use in Database Exploitation

Capacity of the information search depends not only on the resource capacity of the system or searching algorithm but also operative and processing time on users' computer (users).

From the access, study and formation of abbreviation database, we use abbreviations in practical works. We informed a measure by developing a generation function for abbreviations (abbreviations) to apply into the rebuilding of database (database) upon the customer information at Switchboard 108 VNPT Da Nang. We also apply the practicality of the measure like abbreviations indicating sections, and the short insert of the word abbreviation at abbreviation search does bring practical benefits for Switchboard 108 VNPT in information search among customers [5].

## 4 Conclusion

Approach and study abbreviations, aggregate the principles of abbreviation generation, build database of abbreviations to serve users in exploitation, storage, statistics and use; especially the abbreviation use in the formation of indicating sections for better search

of specialized database would be beneficial in enhancing productivity and practical use of data.

Besides, use abbreviations coherently and universally to standardize the system of abbreviations for users, gradually enlarge the system of vocabulary, contribute to the development of language. The proposal of rules, methods in management, building of an abundant storage, convenient exploitation and use, easy update, formation of forum, new addition of abbreviations, etc. are necessary and beneficial.

We continue to expand the storage of abbreviations in many fields, increase the number of auto-updated abbreviations, evaluation of frequent occurrences and abbreviation use; enhance the transfer into many different languages; and expand the searching capacity in multi-language like Vietnamese-Kinh, language of ethnic minorities (Cham, Ede, Thai, Kh'me, etc.), English, French, Chinese, etc. This is a righteously oriented pathway to satisfy a common interest.

**Acknowledgment.** Author group sincerely send our gratefulness to staffs of Switchboard 1080 VNPT Da Nang to create favorable conditions during the process of approach, building of database of abbreviations and abbreviation use, also to exploit and contribute actively for the completion of database.

## References

1. Can, N.T.: Vietnamese Grammar. Publishing House of University and Professional Secondary School, Hanoi (1981)
2. Hang, L.M.: Nom in context of regional culture. In: International Conference About Nom, Between 12–13 November 2004, National Library of Vietnam (2004)
3. Nhan, N.T., Viet, N.T., Nom Na group: Nom Na process. In: Summer Conference 2002 at Maine University (2002)
4. Khanh, P.H., Tuy, N.N.: Study to build up database of abbreviations in service 1080 of Da Nang Post office. In: Summary Record of National Scientific Conference “Some selected issues of information technology and media” (2006)
5. Tuy, N.N., Khanh, P.H.: Abbreviation use in service exploitation of switchboard 108 VNPT Da Nang City. *IJSET Int. J. Innovative Sci Eng. Technol.* **3**(1), 222–227 (2016)
6. Khanh, P.H.: Build database of multi-language vocabulary in form of document RTF Winword. In: Summary Record of National Scientific Conference ICT, rda 2003, pp. 103–110 (2003)
7. Khanh, P.H.: Use programming tools macro VBA, build up text processing facilities. In: Summary Record of the Third Scientific Conference, Da Nang University, pp. 255–261, November 2004
8. Viet, N.T., Bang, D.K.: Terms in Telecommunication abbreviations. Publisher of post office (1999)
9. Thuy, N.T.T.: Vietnamese vocabulary. Remote training curriculum of Can Tho University
10. Van Be, C.: Vietnamese grammar. Remote training curriculum of Can Tho University
11. Thuy, N.T.T., Chinh, N.H.: Overview of language and linguistics. Remote training curriculum of Can Tho University
12. Phap, H.C., Van Hue, N.: Study, collect and build up database of abbreviations in Vietnamese. *J. Sci. Technol. Da Nang Univ.* **7**(80) (2014)

13. Hiep, H.: Build up searching tools by PHP and MySQL. J. Posts Telecommun. Inf. Technol., series 2, September 2004
14. Zahariev, M.: Acronyms. Simon Fraser University, June 2004
15. <http://chuvietnhanh.sourceforge.net/>
16. <http://www.acronymfinder.com>