MÔ HÌNH DỮ LIỆU TỪ VỰNG CỦA TỪ ĐIỂN TIN HỌC TIẾNG ĐỊA PHƯƠNG NGHỆ-TĨNH

Phan Huy Khánh

Đại học Đà Nẵng

1 Vấn đề cơ sở dữ liệu từ vựng và tiếng địa phương

Trong lĩnh vực nghiên cứu ứng dụng tin học xử lý ngôn ngữ tự nhiên, người ta phải xây dựng và tích luỹ các cơ sở dữ liệu (CSDL) từ vựng (lexical database) để từ đó khai thác nhờ các từ điển chuyên dụng khác nhau. Đặc điểm chung của các CSDL từ vựng là nguồn dữ liệu rất lớn, không cùng cách tổ chức và không cùng cách biểu diễn bên trong máy tính. Việc bảo trì, cập nhật và khai thác thường gặp rất nhiều khó khăn. Một trong những nguyên nhân là các nguồn dữ liệu từ vựng lấy từ nhiều nơi, từ các từ điển giấy, hoặc từ internet, không đồng nhất về cách tổ chức, không hoàn toàn giống nhau về nội dung. Lấy ví dụ các từ điển tiếng Việt, mỗi tác giả có một cách riêng để tổ chức và diễn giải các mục từ (entry/headword), nhiều khi rất khác nhau về quan niệm, về thuật ngữ.

Để có được những từ điển phù hợp với nhu cầu sử dụng khác nhau trong máy tính, khi ngày nay các dịch vụ mạng, internet được phổ cập rộng rãi, cần có giải pháp tổ chức phù hợp cho các nguồn dữ liệu từ vựng. Trong các phương pháp phân tích và thiết kế các hệ thống thông tin (cấu trúc, hay hướng đối tượng), để có được đối tượng xử lý là các CSDL vật lý, cần xây dựng mô hình ý niệm dữ liệu (data conceptual model), trước khi chuyển đổi về một mô hình lôgic dữ liệu (data logical model). Đây là giai đoạn quan trọng mang tính quyết định chất lượng của một hệ thống thông tin. Vì vậy cần có một mô hình ý niệm dữ liệu khi xây dựng một CSDL từ vựng.

Hiện nay, nhiều từ điển đơn ngữ, đa ngữ về tiếng Việt đã được xây dựng, sử dụng dưới nhiều hình thức như cài đặt tại máy, tra cứu qua mạng [13]... Nhờ các phương tiện tin học, có thể dễ dàng sưu tập và tích luỹ nguồn dữ liệu từ vựng phong phú này để có được những từ điển chuyên dụng để xử lý tiếng Việt, tuy nhiên vẫn chưa có những từ điển tin học về tiếng địa phương. Như [1] đã chỉ ra, nghiên cứu tiếng địa phương (hay phương ngôn, phương ngữ) không những giúp ích cho việc chuẩn hóa, dạy-học và làm phong phú tiếng Việt, mà còn giúp thực hiện các công trình nghiên cứu tiếng địa phương.

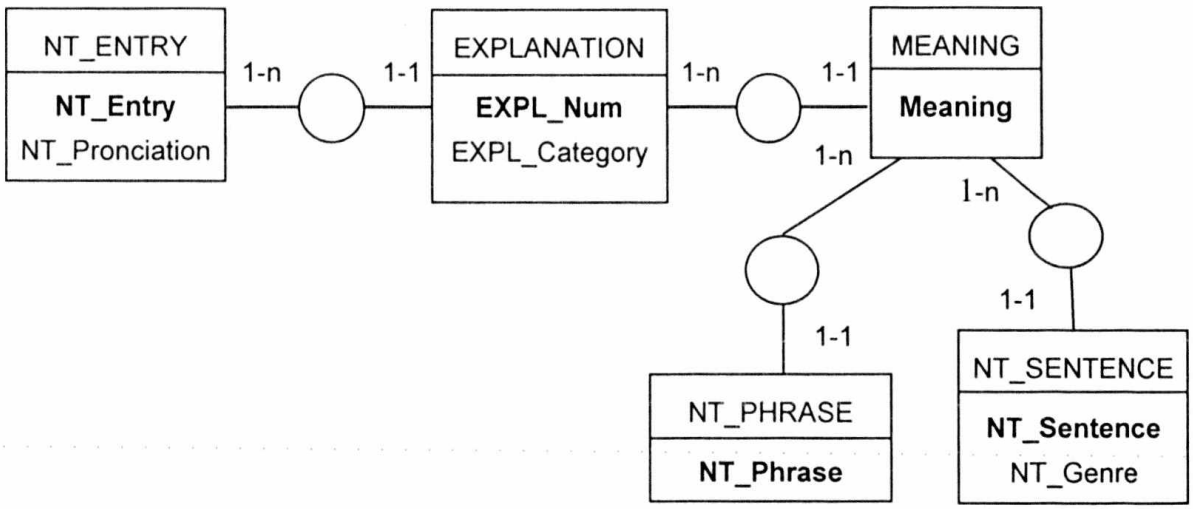
Trong bài báo này, chúng tôi đề xuất giải pháp xây dựng một mô hình ý niệm dữ liệu để từ đó tạo nguồn dữ liệu từ vựng cho từ điển tiếng địa phương Nghệ-Tĩnh (TĐPNT) có tên là Nghê-Tinh Dialectal Dictionary. Chúng tôi đã chọn mô hình thực thể-kết hợp (entity-association model) theo phương pháp phân tích cấu trúc. Chúng tôi đã chọn tiếng địa phương Nghệ-Tĩnh như là ví dụ mẫu đầu tiên minh hoạ quá trình thiết kế hệ thống từ mô hình dữ liệu đã xây dựng. Các tiếng địa phương Việt

Nam khác như Bình-Trị-Thiên-Huế, xứ Quảng, Nam Trung Bộ, Nam Bộ sẽ tiếp tục được đưa vào một CSDL từ vựng lớn hơn cũng từ mô hình này. Riêng những vấn đề về phát âm theo đúng giọng địa phương chưa được giải quyết trong phạm vi bài báo.

2. Xây dựng mô hình dữ liệu từ vựng

2.1. Mô hình ý niệm dữ liệu

Dựa theo cấu trúc của một số từ điển tiếng Việt (8, 9, 10, 11, 12) và [1], từ điển tin học TĐPNT là một tập hợp các mục từ. Mỗi mục từ được phiên theo cách viết (phục vụ phát âm) và có từ một đến nhiều cách giải nghĩa. Mỗi cách giải nghĩa tương ứng với một từ loại (word-category) và có từ một đến nhiều nghĩa phổ thông (popular meaning). Mỗi nghĩa phổ thông có thể có ví dụ : một hoặc nhiều cụm từ, thành ngữ (phrase), câu (sentence) được trích ra từ 5 thể loại : ca dao, hát giặm, hát phường vải, hát ví và hò-vè Nghệ-Tĩnh. Để đơn giản, các mục từ đồng âm nhưng khác cách giải nghĩa trong [1] đều chỉ được xem là một mục từ. Mặt khác, mỗi nghĩa phổ thông xuất hiện trong CSDL được quy ước là "duy nhất" (dãy ký tự có mặt một lần). Các từ ngữ phổ thông không đưa vào làm mục từ trong từ điển.



Hình 1. Mô hình thực thể-kết hợp cho từ điển tin học TĐPNT.

Mô hình có 5 thực thể: mục từ (NT_ENTRY), cách giải nghĩa (EXPLANATION), nghĩa phổ thông (MEANING), cụm từ (NT_PHRASE) và câu (NT_SENTENCE). Mỗi thực thể có một khóa là thuộc tính được in đậm, ví dụ NT_Entry. Các kết hợp giữa các thực thể đều là phân cấp, có hai cặp bản số là (1-n) và (1-1). Chẳng hạn, kết hợp giữa hai thực thể NT_ENTRY và EXPLANATION được hiểu là : mỗi mục từ có tối thiểu 1 và có tối đa n>1 cách giải nghĩa, mỗi cách giải nghĩa chỉ thuộc về 1 và chỉ 1 mục từ. Trong hình 1, mỗi thực thể là một hình chữ nhật, mỗi kết hợp phân cấp là một hình ôvan nối với thực thể bằng các đoạn thẳng ghi bản số (cặp các số nguyên 0, 1, n).

Ví dụ mục từ **bưng** (trong [1] có 4 mục từ riêng) có bốn cách giải nghĩa dựa theo mô hình ý niệm như sau :

1. **bưng** là *động từ*, có một nghĩa phổ thông là *mưng* (nói về mụn nhọt hay vết thương sưng to, phát sốt). Ví dụ *bưng mủ.*

2. **bưng** là *danh từ*, có một nghĩa phổ thông là *tấm che.*

Ví dụ, vè Nghệ-Tĩnh :

Lấy tơi nón che sương

Đất làm bưng che gió

3. **bưng** là *động từ*, có một nghĩa phổ thông là *che lại, bịt lại.*

Ví dụ, hát giặm Nghệ-Tĩnh :

Gánh một gánh đất

Vắt được ba trăm cái nồi

Đập một con đồi

Bưng được mười cái trống

4. **bưng** là *động từ*, có hai nghĩa phổ thông.

- *bê* (dùng tay nâng hay bê lên). Ví dụ, hát giặm Nghệ-Tĩnh :

Cỗ bàn rập rình

Bưng ra hai dãy

Ví dụ khác, hát giặm Nghệ-Tĩnh :

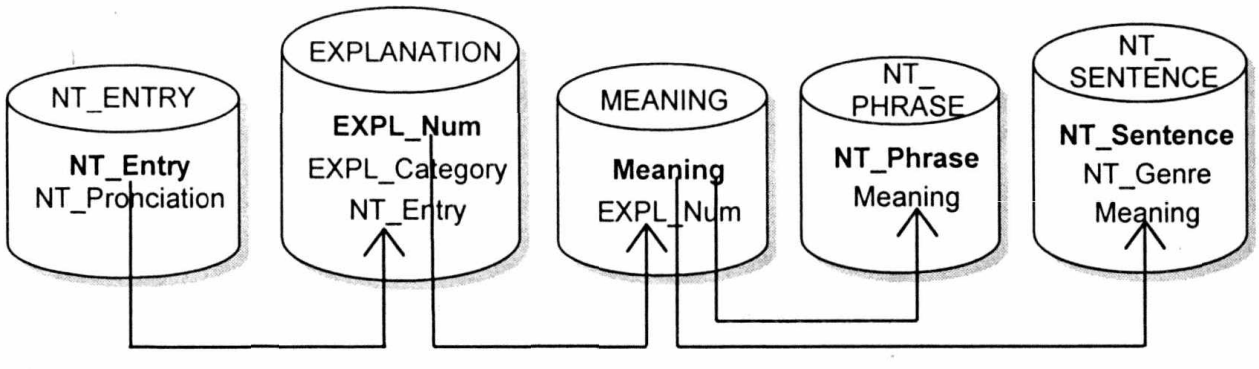
Cỗ năm một bưng ra

- *khiêng.* Ví dụ hát giặm Nghệ-Tĩnh :

Hòn đất to bưng mà nỏ nổi

2.2. Mô hình lôgich dữ liệu

Bước tiếp theo, chuyển mô hình ý niệm dữ liệu về mô hình lôgic dữ liệu, dạng các bảng dữ liệu và mối quan hệ (liên kết lôgic) giữa chúng (có thể biểu diễn bởi lược đồ các quan hệ).



Hình 2. Mô hình logic dữ liệu cho từ điển tin học TĐPNT.

Cách chuyển đổi được thực hiện như sau : mỗi thực thể của mô hình ý niệm dữ liệu trở thành một bảng dữ liệu có cùng khóa với thực thể. Kết hợp giữa hai thực thể trở thành quan hệ giữa hai bảng bằng cách đặt thêm vào bảng "con" (phía bản số 1-1) khóa "ngoại" là khóa của bảng "cha" (phía bản số 1-n). Chẳng hạn thực thể EXPLANATION thành bảng EXPLANATION có khóa ngoại là NT_Entry. Mối quan hệ một-nhiều giữa các bảng là các đường mũi tên trong hình 2 với quy ước chiều đi từ một đến nhiều.

Từ đây, dữ liệu từ vựng được cập nhật trực tiếp vào các bảng để nhận được các tệp CSDL vật lý ở một trong các dạng quen thuộc trong Windows, như Access MDB, FoxPro DBF, hay Excel XLS. Để cập nhật dữ liệu được thuận tiện và sử dụng hệ thống khai thác từ điển đã có (tham khảo [2, 4, 5, 6, 7]), chúng tôi đã sử dụng mẫu văn bản Winword (document template) để làm mô hình logic dữ liệu. Mô hình biểu diễn CSDL từ vựng của từ điển tin học TĐPNT có đạng tổng quát như sau :

Cấu trúc mẫu văn bản	*Giải thích*
NT_Entry	Mục từ
NT_Pronciation	Phiên cách viết của mục từ (để phát âm)
EXPL_Num_1	Cách giải nghĩa 1
EXPL_Category_1	Từ loại
Meaning_1.1	Nghĩa phổ thông 1 cho cách giải nghĩa 1
NT_Phrase_1.1.1	Cụm từ 1 ví dụ cho nghĩa PT 1
...	...
NT_Phrase_1.1.K	Cụm từ thứ K, K≥ 0, ví dụ cho nghĩa PT 1
NT_Genre_1.1.1	Thể loại ví dụ cho nghĩa PT 1
NT_Sentence_1.1.1	Câu tương ứng với thể loại 1
...	...
NT_Genre_1.1.L	Thể loại thứ L, L≥ 0, ví dụ cho nghĩa PT 1
NT_ Sentence_1.1.L	Câu tương ứng với thể loại L
...	...
Meaning_1.M	Nghĩa phổ thông M, M≥ 1, cách giải nghĩa 1
...	...
EXPL_Num_N	Cách giải nghĩa thứ N, N≥ 1

Hình 3. Mẫu văn bản Winword của từ điển tin học TĐPNT.

Mẫu văn bản gồm các dạng thức (style). Mỗi dạng thức thể hiện cách định dạng (format) một đoạn văn bản (paragraph) là cách sử dụng phông chữ (font) trong đoạn và thể thức trình bày đoạn. Hình 4 dưới đây minh hoạ nội dung của mục từ bưng.

Ví dụ mục từ ***bưng***	*Tên dạng thức tương ứng*
bưng	**NT_Entry**
BUWNG	*NT_Pronciation*
1	**EXPL_Num**
động từ	*EXPL_Category*
mưng (mụn nhọt hay vết thương sưng to, phát)	Meaning
bưng mủ	NT_Phrase
2	**EXPL_Num**
danh từ	*EXPL_Category*
tấm che	
vè Nghệ Tĩnh :	***NT_Genre***
Lấy tơi nón che sương *Đất làm bưng che gió*	*NT_Sentence*
3	**EXPL_Num**
động từ	*EXPL_Category*
che lại, bịt lại	Meaning
hát giặm Nghệ Tĩnh :	***NT_Genre***
Gánh một gánh đất *Vắt được ba trăm cái nồi* *Đập một con đồi* *Bưng được mười cái trống*	*NT_Sentence*
4	**EXPL_Num**
động từ	*EXPL_Category*
dùng tay nâng hay bê lên	Meaning
hát giặm Nghệ Tĩnh :	***NT_Genre***
Cỗ bàn rập rình *Bưng ra hai dãy*	*NT_Sentence*
hát giặm Nghệ Tĩnh :	***NT_Genre***
Cỗ năm một bưng ra	*NT_Sentence*
khiêng	Meaning
hát giặm Nghệ Tĩnh :	***NT_Genre***
Hòn đất to bưng mà nỏ nổi	*NT_Sentence*

Hình 4. Ví dụ mục từ bưng của từ điển tin học TĐPNT.

Từ điển tiếng địa phương Nghệ-Tĩnh [1] có tất cả 5901 đơn vị mục từ được sắp xếp theo thứ tự của 30 chữ cái : A(53), Ă(43), Â(11), B(528), C(891), D(312), Đ(395), DZ(54), E(12), Ê(7), G(238), H(204), I(12), K(238), L(416), M(437), N(678), O(36), Ô(29), Ơ(3), P(66), Q(40), R(273), S(196), T(397), TL(15), U(14), Ư(22), V(168), X(113). Các con số trong cặp dấu ngoặc đứng sau mỗi chữ cái là số lượng mục từ tương ứng.

Từ mô hình lôgic dữ liệu, xây dựng được một sơ đồ XML (eXtensible Markup Language) [3, 8, 13] bằng cách sử dụng lại tên các dạng thức trong mẫu văn bản Winword trên đây cho các thẻ (tag) XML.

```
<?xml version="1.0" ?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/TR/WD-xsl" xmlns="http://www.w3.org/TR/REC-html40"
  result-ns="" />
<!DOCTYPE dictionary SYSTEM "tddpnt">
<dictionary name=" TDDPNT" source-language="en" target-language="en,vn">
<dictionary>
...
    <NT_Entry> bưng
        <NT_Pronciation> /BUWNG/ </NT_Pronciation>
            <EXPL_Num> 1 </EXPL_Num>
                <EXPL_Category> động từ </EXPL_Category>
                    <Meaning> mưng (mụn nhọt hay vết thương sưng to, phát) </Meaning>
                        <NT_Phrase> bưng mủ </NT_Phrase>
            <EXPL_Num> 2 </EXPL_Num>
                <EXPL_Category> danh từ </EXPL_Category>
                    <Meaning> tấm che </Meaning>
                        <NT_Genre> vè Nghệ Tĩnh : </NT_Genre>
                        <NT_ Sentence>      Lấy tơi nón che sương
                                                Đất làm bưng che gió </NT_ Sentence>
        ...
    </NT_Entry>
...
</dictionary>
```

Hình 5. Sơ đồ XML tổ chức dữ liệu cho từ điển tin học TĐPNT.

Trong sơ đồ, sau phần tiêu đề khai báo phiên bản của XML và một số khai báo tuỳ chọn khác, là khai báo cấu trúc của từ điển gồm các khai báo mục từ nằm giữa cặp thẻ là <dictionary> và </dictionary>. Mỗi mục từ, nằm giữa cặp thẻ <NT_Entry> và </NT_Entry>, là một tổ hợp các phần tử XML tương ứng với các đoạn của mẫu văn bản biểu diễn cấu trúc lôgic của từ điển tin học TĐPNT. Ví dụ phần tử <EXPL_Category> </EXPL_Category>, v. v

Từ cách biểu diễn này, ta nhận được các tệp XML có tên tệp (filename) chứa phần mở rộng là XML. Để gọi được trình duyệt duyệt xem các tệp XML, cần xây dựng tệp định nghĩa kiểu văn bản DTD (Document Type Definition) và tệp định nghĩa kiểu trình bày CSS (Cascade Style Sheet).

2.3. Chọn bộ mã tiếng Việt

Dữ liệu của từ điển tin học TĐPNT là tiếng Việt nên cần phải chọn một bộ mã để biểu diễn. Cho đến nay đã có nhiều bộ mã tiếng Việt khác nhau được xây dựng và được sử dụng quen thuộc ở Việt nam như TCVN3-ABC, Vietware, VNI, BK TPHCM Hầu hết các bộ mã này đều được xây dựng trên bộ mã ASCII[1] mở rộng, sử dụng 128 vị trí sau bảng, từ 129 đến 256, theo phương pháp "dựng sẵn" (mã hoá cả 134 chữ Việt viết hoa, viết thường, ghép nguyên âm và dấu thanh). Vì chưa có một bộ mã tiếng Việt thống nhất[2] , việc trao đổi tìm kiếm thông tin trong máy tính, trên các trang Web, gặp nhiều khó khăn và phiền phức. Giải pháp trung gian của

---

[1] Hầu hết các bộ mã tiếng Việt hiện nay khác nhau về số bai (byte) sử dụng (1 bai hoặc 2 bai), về cách sắp xếp thứ tự các dấu thanh, và về cách bố trí các chữ Việt có dấu (dựng sẵn) trong bộ mã...

[2] Nhiều chuyên gia đề nghị sử dụng Unicode để thống nhất tất cả các bộ mã tiếng Việt.

chúng tôi là sử dụng một bộ mã trục (pivot code) để chuyển đổi qua lại dễ dàng giữa các bộ mã. Telex được chọn làm mã trục do telex chỉ sử dụng các ký tự ASCII và quen thuộc với nhiều người. Ví dụ chuyển từ TCVN3-ABC qua telex và từ telex qua Unicode, v.v Mã telex đã được chọn để biểu diễn dữ liệu từ vựng tiếng Việt của từ điển [4, 5].

2.4. Nhập nguồn dữ liệu cho từ điển

Nguồn dữ liệu cho từ điển tin học TĐPNT chủ yếu được lấy từ [1]. Chúng tôi đã sử dụng phương pháp đánh dấu quy ước cho trong bảng 6 dưới đây để thao tác cập nhật được dễ dàng và tăng được tốc độ nhập dữ liệu cho nguồn:

Dãy ký tự	*Vị trí*	*Kiểu đoạn (style)*	*Ví dụ gõ vào*	*Kết quả sau khi xử lý*
@	đầu đoạn	NT_Entry	@bưng	**bưng**
n	đầu đoạn	EXPL_Num	1	**1**
Space/Tab	đầu đoạn	EXPL_Category	dt	*động từ*
.	đầu đoạn	NT_Phrase	.bưng mủ	bưng mủ
.k	đầu đoạn	NT_Genre	.5	***vè Nghệ Tĩnh :***
/	cuối đoạn	NT_Sentence	Lấy tơi nón che sương/Đất làm bưng che gió	*Lấy tơi nón che sương* *Đất làm bưng che gió*

Hình 6. Bảng đánh dấu quy ước nhập dữ liệu.

Người sử dụng (NSD) nhập dữ liệu tiếng Việt bằng phương pháp telex trên một trình soạn thảo văn bản tùy ý, không nhất thiết định dạng (như NotePad, hoặc NC Editor), hoặc nhập trực tiếp trên các trang văn bản Winword theo mẫu văn bản đã xây dựng trên đây. Kiểu đoạn của mỗi đoạn được xác định bởi đặt thêm một dãy ký tự quy ước tương ứng cho trong bảng, hoặc ở vị trí đầu đoạn, hoặc ở cuối đoạn, rồi kết thúc bởi phím Enter (tương đương với ký hiệu paragraph-mark ¶). Giá trị của **n = 1, 2** cho biết đó là cách giải nghĩa thứ mấy trong mục từ. Các từ loại được quy ước viết tắt như sau :

d	danh từ	p	phụ từ, hay tổ hợp phụ từ
dg	động từ	k	kết từ, hay tổ hợp kết từ
t	tính từ	tr	trợ từ, hay tổ hợp trợ từ
d	đại từ	th	thán (cảm) từ, hay tổ hợp thán từ

Giá trị **k = 1...5** được quy ước lần lượt là ca dao, hát giặm, hát phường vải, hát ví và hát vè Nghệ-Tĩnh. Ký hiệu / để ngắt dòng các câu trích đoạn ví dụ tương ứng với thể loại **k**, tương đương với ký hiệu ↵ (manual-line-break). Các đoạn nghĩa phổ thông được gõ bình thường, không cần đặt các ký tự quy ước. Cuối cùng, NSD cũng không cần gõ phần phiên theo cách viết theo kiểu gõ telex vì sẽ được tạo ra một cách tự động. Toàn bộ dữ liệu được lưu trữ trong 21 tệp có tên là chữ cái đầu của các mục từ trong tệp lần lượt là A, B, C, D, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V và X. Như vậy, CSDL nguồn cho từ điển tin học TĐPNT đã được tạo xong gồm các tệp văn bản đánh dấu quy ước ở dạng mã ASCII. Ví dụ một đoạn mã tương ứng với mục từ **bưng**:

```
@buwng
1
    dg
muwng (mujn nhojt hay veest thuowng suwng to, phast)
.buwng mur
2
    d
taasm che
.4
Laasy towi nosn che suwowng/DDaast lafm buwng che gios
3
    dg
che laji, bijt laji
.2
Gasnh moojt gasnh ddaast/Vawst dduwowjc ba trawm casi noofi/DDaajp moojt con ddoofi/Buwng
dduwowjc muwowfi casi troosng
4
    dg
dufng tay naang hay bee leen
.2
Coox bafn raajp rifnh/Buwng ra hai daxy
.2
Coox nawm moojt buwng ra
khieeng
.2
Hofn ddaast to buwng maf nor noori
```

Hình 7. Đoạn dữ liệu nguồn đánh dấu quy ước của mục từ bưng trong mã telex.

Từ các tệp dữ liệu nguồn đánh dấu quy ước, dễ dàng viết các thủ tục bằng Macro VBA để chuyển chúng sang văn bản Winword DOC/RTF nhờ một thuật toán tổng quát ở hình 8. Nếu nguồn dữ liệu tiếng Việt đã ở dạng mã telex, có thể sử dụng trình chuyển mã của UniKey, hoặc VietKey, v.v để chuyển từ mã telex thành mã TCVN3-ABC, Unicode, hoặc chuyển sang một bộ mã nào đó mong muốn, trước khi chuyển sang văn bản Winword.

```
Thuật toán 1 : Convert_SrcText_to_Winword_Document
Khởi tạo các biến làm việc trung gian
Xác định các tệp nguồn
Do While Chưa hết tệp nguồn
    Mở một tệp nguồn
    Xác định các tham biến tìm kiếm/thay thế
    Xác định kiểu đoạn cần thay thế : Replacement.Style = NT_Entry
    Do  With Selection.Find
            .Text = Dãy ký tự đánh dấu quy ước, chẳng hạn "@", ".k"...
            .Replacement.Text = ""
            .Forward = True
            .Wrap = wdFindContinue
            .Format = True
        End With
        Selection.Find.Execute Replace := wdReplaceAll
    Loop Until Xử lý hết các dãy ký tự đánh dấu quy ước
Loop    ' Hết tệp nguồn
Kết thúc
```

Hình 8. Thuật toán chuyển nguồn sang văn bản Winword.

Tuy nhiên, do việc chuyển mã không quá phức tạp nên chúng tôi đã xây dựng thuật toán 2 (Convert_TelexCode_to) để giải quyết vấn đề. Các bước xử lý tương tự thuật toán 1, là xây dựng một vòng lặp tìm kiếm các đoạn mã telex của mỗi chữ Việt có dấu (nguồn) trong tệp văn bản đang mở để thay thế bởi mã (đích) tương ứng. Đầu tiên là xử lý (tìm kiếm/thay thế) các đoạn mã telex có độ dài 3, chẳng hạn aaf/ầ, aar/ẩ , sau đó xử lý các đoạn mã telex có độ dài 2, chẳng hạn aa/â, aw/ă Trong trường hợp cần tạo nguồn từ dữ liệu tiếng Việt không ở mã telex, thuật toán 3 (Convert_to_TelexCode) thực hiện chuyển từ mã hiện hành thành mã telex. Có thể minh hoạ quá trình chuyển dữ liệu nguồn đánh dấu quy ước sang các tệp văn bản Winword trong một mã đích nào đó, chẳng hạn TCVN3-ABC, trong thuật toán sau :

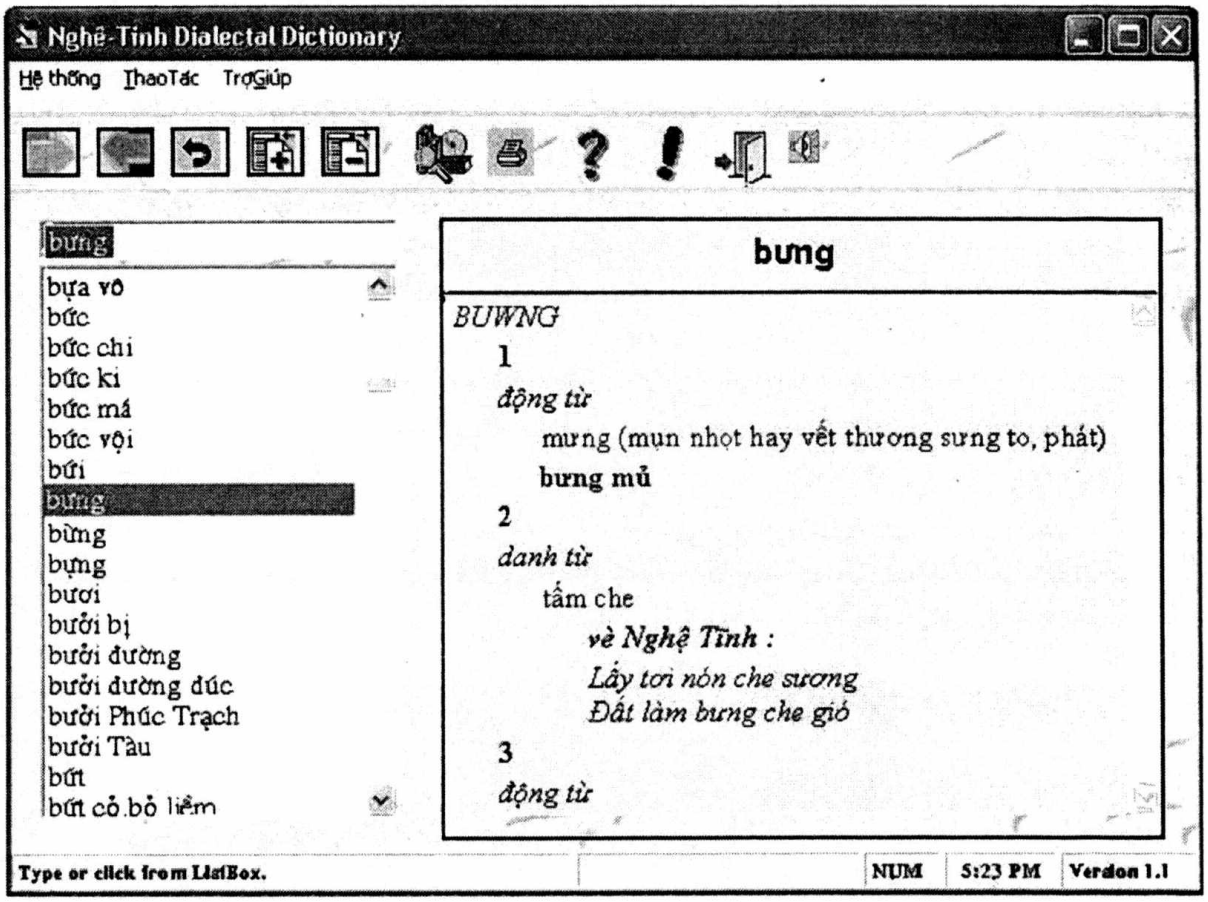
```
Xác định các tệp nguồn đánh dấu quy ước
If      Mã nguồn là telex
Then    Convert_TelexCode_to   ' Xử lý chuyển mã từ telex sang TCVN3-ABC
Else    Convert_to_TelexCode   ' Xử lý chuyển mã từ TCVN3-ABC sang telex
End If
Xác định lại các tệp nguồn có mã là TCVN3-ABC
Convert_SrcText_to_Winword_Document
```

Hình 9. Thuật toán chuyển mã.

Sau khi chuyển nguồn đánh dấu quy ước và chuyển mã, chúng tôi nhận được CSDL từ vựng của từ điển tin học TĐPNT dưới dạng các tệp văn bản Winword DOC/RTF. Từ đây có thể in ra giấy thành từ điển tra cứu theo mẫu in tùy ý, hoặc chuyển sang HTML/XML để sử dụng các trình duyệt [5, 6], hay cài đặt trên CD-ROM.

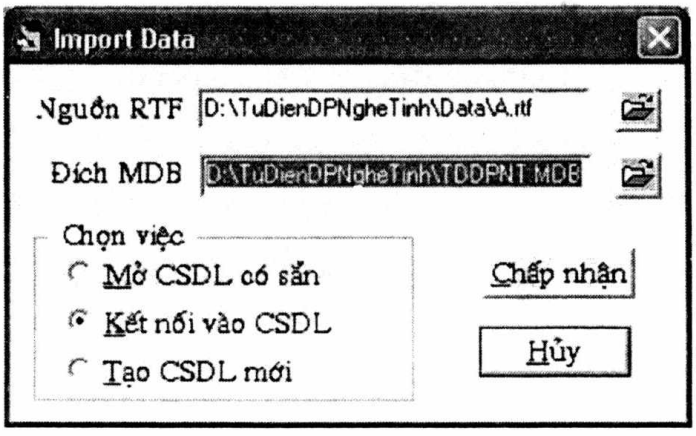
3. Xây dựng từ điển tin học TĐPNT

Cửa sổ làm việc chính của từ điển tin học TĐPNT Nghê-Tinh Dialectal Dictionary gồm 4 vùng. Vùng 1 có thanh tiêu đề ở trên cùng và thanh trạng thái ở dưới cùng của cửa sổ. Vùng 2 gồm các lệnh HệThống, ThaoTác, và TrợGiúp. Vùng 3 gồm các nút lệnh dễ nhận biết sử dụng. Vùng 4 để tra cứu từ điển, gồm cột danh sách các mục từ bên trái và nội dung tương ứng ở cột bên phải. Hệ thống tra cứu từ điển TĐPNT được phát triển từ mã nguồn mở của hệ thống khai thác CSDL từ vựng đa ngữ [5, 6, 7]. Nguyên tắc hoạt động của hệ thống như sau : các tệp dữ liệu văn bản DOC được chuyển đổi thành RTF trước khi chuyển đổi sang CSDL trung gian Access MDB nhờ các lệnh chuyển (data import). Tiếp theo, hệ thống đưa kết quả lên màn hình để phục vụ tra cứu.



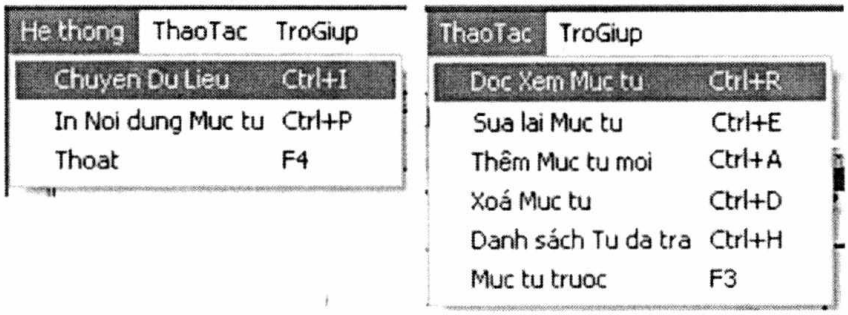
Hình 10. Giao diện chính của từ điển tin học TĐPNT.

Các thao tác như sau : khi sử dụng Nghê-Tinh Dialectal Dictionary lần đầu tiên, hệ thống đưa ra lời nhắc NSD cần đọc CSDL từ vựng (giai đoạn import data to dictionary) từ các tệp văn bản Winword RTF để chuyển thành CSDL trung gian Access MDB. Bằng cách gọi lệnh đơn Chuyển dữ liệu, chọn mục việc Tạo CSDL mới, khi đó, lần lượt các tệp văn bản RTF nguồn được chuyển tải vào từ điển. Những lần chạy chương trình sau này, hệ thống mặc nhiên sử dụng CSDL trung gian MDB đã có sẵn. Tuy nhiên, NSD vẫn có thể chọn lệnh bổ sung dữ liệu mới từ một tệp văn bản nguồn RTF hợp lệ (lệnh Kết nối vào CSDL), hoặc mở lại CSDL MDB đã có (lệnh Mở CSDL có sẵn) tuỳ theo yêu cầu. Hình 11 minh họa hộp thoại của lệnh HệThống-Chuyển dữ liệu.



Hình 11. Hộp thoại lệnh đơn đọc dữ liệu nguồn vào từ điển.

Khi từ điển đã có dữ liệu và sẵn sàng làm việc, NSD tìm chọn để đọc-xem một mục từ từ danh sách các mục từ đã được sắp xếp theo thứ tự chữ cái (xem mục 2.1 trên đây) và dấu thanh : không dấu, huyền, ngã, hỏi, sắc, nặng. NSD có thể sao chép mục từ và in ra giấy nội dung mục từ đó, hay có thể sửa lại mục từ nhờ lệnh ThaoTác-Sửa lại Mục từ (phím tắt ^E). Hệ thống cho phép NSD tìm xem lại một mục từ hoặc các mục từ đã tra trước đó từ một danh sách, hoặc thêm một mục từ mới, hoặc xoá bỏ một mục từ.



Hình 12. Một số lệnh đơn của từ điển tin học TĐPNT.

Khi chạy chương trình, NSD nhắp chuột tại nút đọc trên thanh công cụ để nghe đọc một mục từ bất kỳ hiện đang tra cứu.

4. Kết luận

Từ điển tin học TĐPNT chạy trong Windows 9x. Đây là một đóng góp của chúng tôi trong quá trình nghiên cứu ứng dụng tin học cho lĩnh vực xử lý ngôn ngữ tự nhiên, xử lý tiếng Việt, góp phần giải quyết từng bước những vấn đề đa ngữ của tiếng Việt đặt ra. Trong bối cảnh này, chúng tôi đã và đang tiếp tục nghiên cứu xử lý tin học về tiếng Việt như xử lý văn bản tiếng ÊĐê, tiếng Chăm, chữ Hán (tiếng Trung quốc, trên cơ sở từ Hán-Việt) và xây dựng các từ điển đơn ngữ, đa ngữ

Từ điển tin học TĐPNT giúp NSD nghe hiểu được lời nói, chữ viết khi tiếp xúc với người Nghệ-Tĩnh, hiểu thêm về lịch sử tiếng Việt, hiểu thêm bản sắc văn hoá địa phương của một vùng đất miền Trung. Từ điển giúp dạy-học môn Tiếng Việt được tốt hơn. Từ kết quả đã có, có thể xây dựng một công cụ kiểm sửa lỗi chính tả, ngữ pháp mà NSD, người Nghệ-Tĩnh, thường mắc phải. Đây cũng là những yếu tố cần thiết để tiếp tục xây dựng các từ điển tin học tiếng địa phương khác trên đất nước Việt Nam.

Với nguồn dữ liệu từ vựng đã có và với khả năng cập nhật, sửa đổi, từ điển có thể tiếp tục được bổ sung mục từ, các trích đoạn từ các thể loại hò, hát giặm, hát phường vải, hát ví, hát vè, hay trích đoạn văn, thơ, ca dao... có sử dụng các từ địa phương Nghệ-Tĩnh. Để từ điển tin học TĐPNT trở thành một sản phẩm hoàn chỉnh, được phát triển và phổ biến sử dụng rộng rãi theo hướng mã nguồn mở, thoả mãn điều kiện mã tiếng Việt đọc được (readability), cần tiếp tục bổ sung các chức năng mới cho Nghê-Tinh Dialectal Dictionary như khả năng tra chéo mục từ (tra cứu một từ nằm trong phần giải nghĩa), tra cứu bằng nhiều phương pháp (click-and-see, autolook), tìm đưa ra các câu nói tiếng địa phương tương đương, v.v

- **Lời cảm ơn**: Bài báo tham khảo kết quả đồ án tốt nghiệp kỹ sư ngành CNTT của em Lê Thị Phương, sinh viên khoá 1998, đã bảo vệ thành công tháng 6/2003: "Xây dựng từ điển địa phương tiếng Nghệ-Tĩnh" do tác giả hướng dẫn trực tiếp, tại khoa CNTT và ĐTVT, trường Đại học Kỹ thuật, Đại học Đà Nẵng. Tác giả chân thành cảm ơn.

Tài liệu tham khảo

1. Nguyễn Nhã Bản, Phan Mậu Cảnh, Hoàng Trọng Canh, Nguyễn Hoài Nguyên, *Từ điển tiếng địa phương Nghệ-Tĩnh*, NXB Văn hóa Thông tin, Hà Nội, 1999, 460tr.
2. Phan Huy Khánh, Thiết kế từ điển phát âm tiếng Việt trong Microsoft Windows, *Tạp chí Khoa học Công nghệ*. Số 19+20, 1999, tr.21-27.
3. Phan Huy Khánh (chủ trì), Thiết kế hệ thống khai thác cơ sở dữ liệu từ vựng đa ngữ Pháp-Anh-Việt, *Đề tài NCKH cấp Bộ, mã số B2001-15-04, Đà Nẵng 2001-2002*, Lưu Bộ Giáo dục và Đào tạo.
4. Phan Huy Khánh, Võ Trung Hùng. Thiết kế cơ sở dữ liệu đa ngữ ngữ pháp tiếng Việt. *Tạp chí Khoa học Công nghệ*, No 36+37, 2002, tr.19-24.
5. Phan Huy Khánh, Xây dựng cơ sở dữ liệu từ vựng đa ngữ sử dụng dạng thức văn bản RTF Winword, *Kỷ yếu Hội thảo Khoa học Quốc gia Lần thứ nhất*, ICT.rda'2003 HàNội, 2003, tr103-110.
6. M. Mangeot-Lerebours, Environnements centralisés et distribués pour lexicographes et lexico-logues en contexte multilingue, *Luận án Tiến sĩ*, 9/2001, UJF, CH Pháp, ĐHTH Joseph Fourrier
7. Hoàng Phê, *Từ điển chính tả*, Trung tâm Từ điển học, NXB Đà Nẵng, 1995, 511tr.
8. Hoàng Phê, *Từ điển tiếng Việt*, Trung tâm Từ điển học, NXB Đà Nẵng, 1997, 1130tr.
9. Nguyễn Kim Thản, *Ngữ pháp tiếng Việt*, NXB Giáo dục, 1997, 232 tr.
10. Nguyễn Như Ý, *Đại từ điển tiếng Việt*, NXB Văn hóa-Thông tin, Hà Nội 1999, 1892 tr.

VNU JOURNAL OF SCIENCE, Nat., Sci., & Tech., T.XX, $N_0$1, 2004

DATA MODELS FOR THE NGHE-TINH DIALECTAL DICTIONARY

Phan Huy Khanh

The University of Danang

In the research of natural language processing (NLP), one must always accumulate and update of more from many of lexical data resources of heterogeneous formats for various applications. These resources are often difficult to maintain and to manipulate. It is necessary reconstructing a specific dictionary for every new application. Following the methods of analysis and design of the information systems, it is necessary to create a data conceptual model and then convert it in a data logical model in order to construct a lexical data base. Currently in Vietnam, there are already some Vietnamise dictionaries on computer, but it doesn't exist more dialectal dictionary.

We present in this paper a solution of constructing of data models in order to create a Nghe-Tinh dialectal dictionary. We construct an entity-association model to represent the relationshif between the entry (headword), explanation, popular meaning, phrase et sentence from a publish paper Nghe-Tinh dialectal dictionary. This model is convert into Winword document format to update the Nghe-Tinh dialectal lexical database in the pivot telex code. By using an open sources of a software system of consulting the multiligual lexical database developed by us at the University of Danang, we have build a first version of Nghe-Tinh dialectal dictionary on computer. The lexical resource of this dictionary contains about 5000 entries with the possibility of update and readable. In the same time, the entity-association model is also converted into Access MDB table and XML format.