NGHIÊN CỨU VÀ XÂY DỰNG TỪ ĐIỂN TIẾNG VIỆT CHO MÁY TÍNH

(Building a Vietnamese Computational Lexicon)

Vũ Xuân Lương
Trung tâm từ điển học Vietlex

Nguyễn Thị Minh Huyền
Trường Đại học Khoa học Tự nhiên Hà Nội

Tóm tắt

Trong xử lí ngôn ngữ tự nhiên (Natural Language Processing), từ điển cho máy tính (Machine Readable Dictionary - MRD) là một dạng tài nguyên thiết yếu cho các bài toán phân tích ngôn ngữ từ đơn giản đến phức tạp. Một kho từ vựng chất lượng tốt phải cung cấp được cho các hệ thống xử lí ngôn ngữ tự nhiên các thông tin ngôn ngữ ở nhiều tầng bậc khác nhau như hình thái, ngữ pháp, ngữ nghĩa, tốt hơn nữa là có thể phục vụ cả các hệ thống xử lí đơn ngữ và đa ngữ. Trong báo cáo này, chúng tôi trình bày việc nghiên cứu và xây dựng Từ điển tiếng Việt dùng cho máy tính (Vietnamese Computational Lexicon – VCL), với mục tiêu đặt ra trước mắt là cung cấp ngữ liệu phục vụ phân tích cú pháp tiếng Việt. Chúng tôi sẽ giới thiệu mô hình ngữ liệu cho VCL, quy trình xây dựng VCL và những vấn đề cần phải tiếp tục nghiên cứu, giải quyết trong tương lai.

1. GIỚI THIỆU

Trên thế giới, việc xây dựng loại từ điển dạng MRD áp dụng trong các ứng dụng xử lí ngôn ngữ tự nhiên là rất phổ biến. Đã có nhiều MRD được xây dựng, cả cho các ứng dụng xử lí đơn ngữ và đa ngữ, với những quan niệm và xuất phát điểm riêng (Nguyen, 2006).

Với các kho từ vựng đơn ngữ, có thể kể đến nhiều dạng từ điển cung cấp các thông tin ở các tầng bậc khác nhau. Chẳng hạn, những dự án như BDLEX, CELEX, MULTEXT xây dựng các kho từ vựng chứa thông tin ở mức ngữ âm, hình thái - cú pháp học cho nhiều thứ tiếng Ấn – Âu. Ở tầng bậc cú pháp, nhiều mô hình từ điển cung cấp các thông tin ngôn ngữ rất phong phú, cả về khả năng kết hợp cú pháp cũng như những ràng buộc ngữ nghĩa hay các chức năng trong các cấu trúc ngữ pháp như GENELEX, EAGLES cho các ngôn ngữ Ấn – Âu, CKIP cho tiếng Trung.

Thiên về ngữ nghĩa, các kho từ vựng dạng *WordNet* tạo ra một tập hợp từ vựng đồ sộ, theo đó các từ được sắp xếp trong dãy của những tập hợp đồng nghĩa, giúp cho việc xác định nghĩa của từ và để phân biệt được nghĩa đang xét với các nghĩa khác. Nguyên lí tổ chức chung của Wordnet là mạng lưới quan hệ ngữ nghĩa. Đó là quan hệ đồng nghĩa (synonymy): *dog – domestic dog*; quan hệ trái nghĩa (antonymy): *rich – poor*; quan hệ trên dưới (hyponymy): *maple – tree*; quan hệ chỉnh thể – bộ phận (meronymy): *body – limb*; quan hệ kéo theo (entailment): *snore – sleep* (cho động từ); v.v. Dạng kho từ vựng này rất hữu ích cho việc gán nhãn ngữ nghĩa cũng như việc truy cập vào ngữ nghĩa của văn bản.

Những năm gần đây, cần phải kể đến sự phát triển của những dự án xây dựng kho từ vựng dạng *FrameNet*, dựa trên ngữ nghĩa học và kho văn bản. Mục đích là đưa ra bằng chứng về khả năng kết hợp ngữ nghĩa và cú pháp của từng từ trong từng nét nghĩa của chúng, với sự giải thích có trợ giúp của máy tính trên các câu ví dụ và được trình bày tự động bằng những bảng kết quả. FrameNet cho tiếng Anh hiện bao gồm 8900 mục từ, trong đó hơn 6100 mục từ được chú giải đầy đủ, trên 625 khung từ vựng và được minh hoạ trong hơn 135.000 câu ví dụ.

Về các kho từ vựng đa ngữ, trước tiên phải nhắc đến dự án đồ sộ *EDR* cho cặp tiếng Anh - Nhật. EDR được thiết kế dựa trên 11 từ điển con, bao gồm: từ điển khái niệm, từ điển đơn ngữ, từ điển song ngữ, v.v. Mỗi từ điển đơn ngữ Anh/Nhật bao gồm các mục từ với các thông tin ngữ pháp dưới dạng danh sách các thuộc tính và có liên kết tới các khái niệm trong từ điển khái niệm. Kho từ vựng này về sau được đánh giá là thiết kế chưa kĩ lưỡng nên hiệu quả khai thác chưa cao. Ra đời sau dự án EDR là nhiều dự án từ điển đa ngữ có quy mô tương đối lớn khác như ISLE / MILE của nhóm EAGLES, các dự án Wordnet đa ngữ, dự án Papillon, v.v.

Với sự phát triển đa dạng của các dự án xây dựng từ điển cho xử lí ngôn ngữ vốn đòi hỏi rất nhiều công sức, các nỗ lực phát triển một chuẩn mô hình từ điển để nâng cao khả năng trao đổi và dùng lại của các từ điển đã được hội tụ vào dự án LMF (ISO, 2008) được khởi động từ năm 2002. Dự án này đưa ra một siêu mô hình từ vựng, trong đó mỗi mục từ được mô tả ở nhiều tầng bậc khác nhau, với các khối thông tin đơn ngữ (ngữ âm, hình thái, cú pháp, ngữ nghĩa) và đa ngữ. Theo mô hình này, việc xây dựng một kho từ vựng có thể được làm dần dần, tập trung theo từng khối thông tin.

Đối với việc xây dựng từ vựng tiếng Việt cho máy tính, ngoài các công trình từ điển được xây dựng cho một số hệ thống dịch máy không được phổ biến và chia sẻ rộng rãi, hiện nay các nhóm nghiên cứu xử lí tiếng Việt mới chỉ có sẵn các kho từ vựng với thông tin từ loại và tiểu từ loại đi kèm (ví dụ công trình của Nguyen et al, 2007), còn các thông tin có khả năng phục vụ cho các phân tích ngôn ngữ mức sâu hơn (cú pháp, ngữ nghĩa, …) thì hầu như không có. Do vậy trong khuôn khổ đề tài KC.01.01/06-10, chúng tôi đặt ra mục tiêu xây dựng một kho từ vựng nhằm phục vụ cho cộng đồng nghiên cứu xử lí tiếng Việt, bước đầu là cung cấp thông tin ngôn ngữ cho xử lí cú pháp tiếng Việt. Mô hình ngữ liệu của kho từ vựng được xây dựng theo chuẩn LMF, nhằm đảm bảo khả năng phát triển tiếp ngữ liệu trong các giai đoạn sau. Trong các phần tiếp theo của bài báo này, chúng tôi sẽ trình bày nội dung, cấu trúc kho ngữ liệu VCL và những vấn đề cần phải tiếp tục nghiên cứu, giải quyết. Chúng tôi hi vọng rằng, VCL sẽ trở thành nguồn tri thức cơ bản về từ vựng tiếng Việt, có thể được áp dụng trong các ứng dụng xử lí ngôn ngữ tự nhiên có liên quan đến tiếng Việt một cách rộng rãi.

2. LỰA CHỌN ĐƠN VỊ TỪ VỰNG

Với mục đích xây dựng một từ điển điện tử về tiếng Việt, cho nên vấn đề đặc điểm của tiếng Việt sẽ được chúng tôi quan tâm hàng đầu. Tuy nhiên, bước đầu chúng tôi chỉ quan tâm đến những vấn đề mà nhu cầu thực tế về xử lí tiếng Việt đang đòi hỏi, các vấn đề khác sẽ không được đề cập trong bài báo này. Chúng tôi xác định từ ngữ được thu thập trong VCL bao gồm:

Từ cơ sở (từ gốc): bao gồm các từ đơn – trong sự đối lập với từ ghép – có hình thức chính tả thuần Việt: *cha, mẹ, nhà, bàn, đi, học, hát, xanh, đỏ*, v.v. Các yếu tố Hán-Việt không hoạt động độc lập (không tự thân là từ), nhưng có khả năng cấu tạo từ lớn cũng thuộc lớp từ này. Ví dụ: *bất* (bất bình đẳng, bất bình thường, bất di bất dịch, ...); *vô* (vô thưởng vô phạt, vô chính phủ, vô căn cứ, ...); *hoá* (công nghiệp hoá, hiện đại hoá, tư sản hoá, ...); *siêu* (siêu nhân, siêu lợi nhuận, siêu liên kết, ...), v.v.

Từ phái sinh: bao gồm các từ ghép – trong sự đối lập với từ đơn – có hình thức chính tả thuần Việt. Nằm trong lớp từ này là tất cả các từ ghép và các từ láy: *đất nước, binh lính, mua bán, học sinh, chuồn chuồn, trong trắng, nhanh nhẹn*, v.v.

Thuật ngữ khoa học – kĩ thuật: bao gồm các thuật ngữ được dùng phổ biến trong đời sống xã hội: *bằng sáng chế, bất đẳng thức, bất bạo động, cách mạng xanh, dây tiếp địa, đạo hàm, hàm số, chấn tử*, v.v.

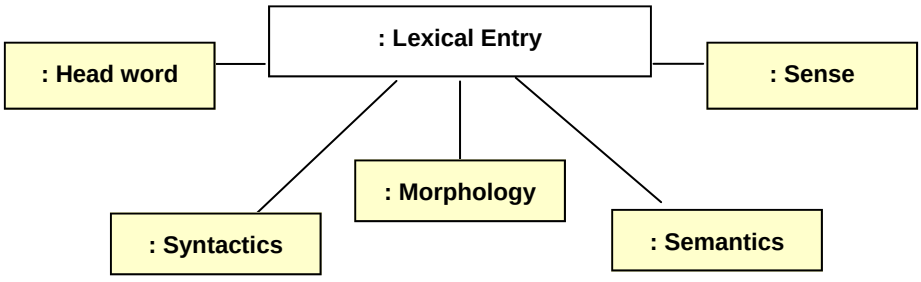
Từ vay mượn: bao gồm các từ mượn có nguồn gốc Ấn – Âu, được thể hiện bằng dạng chính tả phiên âm hoặc giữ nguyên gốc: *vi-ô-lông, a-pa-tít, internet, online, weblog*, v.v.

Từ tắt và kí hiệu: *kg, cm, mg, www, HIV, GDP, VAC*, A, @, X, v.v.

Cách phân loại đơn vị từ vựng như vậy sẽ giúp cho việc chuyển dịch tiếng Việt sang ngôn ngữ khác được thuận lợi hơn. Với hầu hết các từ trong nhóm *từ cơ sở* sẽ có các từ tương đương trong ngôn ngữ khác theo mối tương quan 1 – 1; một số các từ trong nhóm *từ phái sinh* có thể sẽ không có mối tương quan 1 – 1, v.v.

3. XÁC ĐỊNH CẤU TRÚC CHO VCL

Một mục từ của từ điển điện tử thường cung cấp tri thức về chính tả, ngữ âm, từ nguyên, cấu tạo từ, khả năng kết hợp, quan hệ ngữ pháp, quan hệ ngữ nghĩa, v.v. (Vũ Xuân Lương, 2002) của từ ngữ. Những tri thức này tuỳ thuộc vào từng ngôn ngữ và tuỳ thuộc vào từng mục đích sử dụng mà có thể có những yêu cầu thể hiện khác nhau. Nhưng nhìn trên tổng thể, một từ điển như vậy phải được xây dựng dựa trên những nét phổ quát cho mọi ngôn ngữ. Mục đích của phần này là đưa ra lí do lựa chọn mô hình biểu diễn thông tin và cách thức biểu diễn thông tin trong từ điển. Các thông tin mô tả được thể hiện trên 3 bình diện: hình thái học, cú pháp học và ngữ nghĩa học.



Hình 1. Cấu trúc tổng quát của một mục từ.

3.1. Thông tin hình thái (Morphology)

Từ của tiếng Việt, trong cấu tạo, không có căn tố và phụ tố; trong ngữ nghĩa, không có các ý nghĩa thuộc phạm trù hình thái (giống, số, cách); trong hoạt động tạo câu, các mối liên hệ ngữ pháp không biểu hiện ở sự biến hình mà biểu hiện bằng trật tự từ. Vì những lẽ đó, khi xét về tính hình thái của tiếng Việt, thông thường chỉ xét về vấn đề *cấu tạo từ*.

Thông tin về cấu tạo từ khi được kết hợp với thông tin syntactics và semantics sẽ có ích cho các nghiên cứu về tách từ (word segmentation), đoán định đơn vị từ trong văn bản tiếng Việt. Chẳng hạn đoán định cụm từ và từ (*sữa bò* và *bò sữa*, *tấm vải* và *vải tấm*, *xay máy* và *máy xay*, ...), đoán định cơ chế sinh từ láy, v.v. Trong VCL, các dạng cấu tạo từ được chú ý như sau:

- từ đơn: *simple word*
- từ ghép: *composite word*
- từ láy: *reduplicative word*
- từ vay mượn: *borrowed word*
- từ tắt: *abbreviation*
- kí hiệu: *symbol*

```
bàn N
   headWord
        |
        +--written form : bàn
   morphology
        |
        +--word type : simple word
   def : đồ thường làm bằng gỗ, có mặt phẳng và chân đỡ...
```

Hình 2. Thông tin *Morphology* của "bàn".

Thông tin hình thái được mô tả trong VCL chỉ mới dừng lại ở mức gán nhãn bậc một cho mỗi đơn vị từ vựng, các thông tin ở mức sâu hơn chúng tôi chưa có điều kiện đề cập tới.

3.2. Thông tin cú pháp (Syntactics)

Thông tin về loại từ (category)

Các từ thường có chung đặc điểm ngữ pháp và ý nghĩa khái quát, như danh từ, động từ, tính từ, v.v. Mỗi loại từ như vậy phản ánh khả năng kết hợp và chức năng cú pháp khác nhau. Chẳng hạn khi tạo câu, nếu vị ngữ là *danh từ* thì phải dùng *là*, ngược lại nếu vị ngữ là *tính từ* thì không cần *là* (Nguyễn Kim Thản, 1997): *đây là quyển sách; sách này hay quá*. Việc phân định các loại từ là nhằm mục đích tạo câu cho đúng, do vậy việc mô tả chúng là có ý nghĩa. Trong VCL đề cập đến 14 loại sau:

idPOS	vnPOS	enPOS	symbolPOS
1	danh từ	noun	N
2	động từ	verb	V
3	tính từ	adjective	A
4	số từ	numeral	M

5	định từ	determiner	D
6	đại từ	pronoun	P
7	phụ từ	adverb	R
8	giới từ	preposition	O
9	liên từ	conjunction	C
10	trợ từ	auxiliary word	I
11	cảm từ	emotivity word	E
12	yếu tố cấu tạo từ	component stem	S
13	từ tắt	abbreviation	Y
14	không xác định	undetermined	U

Thông tin về tiểu loại từ (subcategory)

Phân định loại từ không những phải đạt yêu cầu khoa học mà còn phải mang tính thực dụng (Nguyễn Kim Thản, 1997). Trong mỗi loại từ như vậy, lại có nhu cầu phân ra thành những tiểu loại nhỏ hơn. Trong VCL đề cập đến 28 loại sau:

idPOS	idSubPOS	vnPOS	enPOS	symbolPOS
1	1	danh từ riêng	proper noun	Np
1	2	danh từ đơn thể	countable noun	Nc
1	3	danh từ tổng thể	collective Noun	Ng
1	4	danh từ chỉ loại	classifier noun	Ns
1	5	danh từ trừu tượng	abstract noun	Na
1	6	danh từ đơn vị	unit noun	Nu
2	7	động từ nội động	intransitifve verb	Vi
2	8	động từ ngoại động	transitive verb	Vt
2	9	động từ trạng thái	state verb	Vs
3	10	tính từ tính chất	property adjective	Ap
3	11	tính từ quan hệ	relative adjective	Ar
3	12	tính từ tượng thanh	onomatopoetic adjective	Ao
3	13	tính từ tượng hình	pictographic adjective	Ai
4	14	số từ số lượng	cardinal numeral	Mc
4	15	số từ thứ tự	ordinal numeral	Mo
5	16	định từ	determiner	D
6	17	đại từ xưng hô	personal pronoun	Pp
6	18	đại từ chỉ định	demonstrative pronoun	Pd
6	19	đại từ số lượng	quality pronoun	Pq
6	20	đại từ nghi vấn	interrogative pronoun	Pi
7	21	phụ từ	adverb	R
8	22	giới từ	preposition	O
9	23	liên từ	conjunction	C

10	24	trợ từ	auxiliary word	I
11	25	cảm từ	emotivity word	E
12	26	yếu tố cấu tạo từ	component stem	S
13	27	từ tắt	abbreviation	Y
14	28	không xác định	undetermined	U

Phân loại từ là một công việc khó khăn và phức tạp. Chúng tôi luôn mong muốn đưa ra được một danh sách từ loại sao cho khi tổng hợp lại sẽ không bỏ sót một trường hợp nào. Nhưng ngôn ngữ là một hiện tượng xã hội đặc biệt, nên rất khó đòi hỏi việc phân loại từ đạt được đầy đủ những yêu cầu theo như mong muốn đó.

Thông tin về mẫu động từ (verb pattern)

Trong tiếng Việt, có hai nhóm thực từ có số lượng lớn và đối lập nhau một cách rõ rệt về ý nghĩa, hình thức thể hiện, đó là *thể từ* (biểu thị thực thể) và *vị từ* (từ làm vị ngữ). Trong vị từ thì động từ đóng một vai trò rất quan trọng. Trong các ngôn ngữ Ấn-Âu, đặc biệt là tiếng Anh và tiếng Pháp, vị ngữ bao giờ cũng là động từ được chia ở những thời và thể nhất định (Nguyễn Minh Thuyết & Nguyễn Văn Hiệp, 2004). Trong tiếng Việt, không phải động từ nào cũng làm vị ngữ. Về vai trò của vị ngữ trong câu, bước đầu chúng tôi chỉ mới quan tâm tới loại động từ, chứ chưa có điều kiện quan tâm tới loại tính từ. Trong VCL, đưa ra 3 mẫu động từ như sau:

Values	**Comment**
Sub+V	động từ không đòi hỏi bổ ngữ: *Chim bay. Bé đang ngủ*
Sub+V+Obj	động từ đòi hỏi một bổ ngữ: *Tôi đọc sách. Nó ngồi xuống sàn.*
Sub+V+Obj+Obj	động từ đòi hỏi hai bổ ngữ: *Tôi tặng hoa cho mẹ. Bà bắt cháu ăn. Họ gọi ông là vị thánh sống.*

```
bàn V
   ...
    syntactics
         |
         +--category : V
         |
         +--subcategory : Vt
         |
         +--verb pattern : Sub+V+Obj
    def : trao đổi ý kiến về việc gì hoặc vấn đề gì.
    exa : bàn kế hoạch ~ bàn chuyện thời sự.
```

Hình 4. Thông tin *Syntactics* của "bàn" với ý nghĩa *động từ*.

3.3. Thông tin ngữ nghĩa (Semantics)

3.3.1. Ràng buộc Logic (logical constraint)

Ý nghĩa phạm trù (categorial meaning)

Các ngôn ngữ có thể có một hệ thống *từ loại ngữ nghĩa* căn bản giống nhau. Có hai loại ngữ nghĩa lớn, một loại *biểu thị thực thể* (thể từ) và một loại *biểu thị thuộc tính của thực thể* hoặc *thuộc tính của thuộc tính* (gọi là thuộc từ - mang ý nghĩa trừu tượng). Đại từ và phần lớn danh từ là thể từ, nhưng cũng có nhiều danh từ là thuộc từ (danh từ chỉ tình cảm, màu sắc, hình dáng, v.v.) (Hoàng Phê, 2008). Trong hai loại lớn lại phân chia ra thành các loại nhỏ, trong mỗi loại nhỏ lại được phân chia ra loại nhỏ hơn. VCL tổ chức *từ loại ngữ nghĩa* theo mô hình quan hệ hình cây, gần 100 tiểu loại. Cây ngữ nghĩa này được tham khảo từ dự án *TCL* (*Thai Computational Lexicon*) (Charoenporn, 2004) có hơn 60.000 mục từ Thái – Anh, được mô tả trên 3 bình diện: hình thái học, cú pháp học và ngữ nghĩa học, v.v...

```
SEMANTIC TREE
  |
  + Thực thể : Concrete Thing
  |   |
  |   + Vật hữu sinh : Living Thing
  |   |    |
  |   |    + Con người    : People
  |   |    + Động vật     : Animal
  |   |    + Vi sinh vật : Microorganism
  |   |    + Thực vật     : Plant
  |   |    ...
  |   + Vật vô sinh : Non Living Thing
  |   |    |
  |   |    + Vật dụng : Artifact
  |   |    ...
  |   + Vị trí : Location
  |   |    ...
  + Trừu tượng : Abstraction
  |   |
  |   + Lĩnh vực tri thức: Field Of Knowledge
  |   + Trạng thái : State
  |   + Hoạt động  : Action
  |   + Quan hệ    : Relation
  |   ...
```

Như vậy, mỗi đơn vị từ vựng trong VCL ngoài việc được gán nhãn từ loại ngữ pháp (học sinh – Nc) còn được gán thêm một nhãn từ loại ngữ nghĩa (học sinh – Person). Việc làm này giúp cho việc phân loại từ được triệt để hơn, hoặc giúp cho việc phân tích cú pháp được sâu sắc hơn.

Từ đồng nghĩa (synonym): Đồng nghĩa là hiện tượng các từ khác nhau về âm thanh nhưng có ý nghĩa giống nhau hoặc gần giống nhau, do đó trong nhiều hoàn cảnh ngôn ngữ cụ thể, chúng có thể thay thế cho nhau được.

Từ trái nghĩa (opposite): Trái nghĩa là hiện tượng các từ khác nhau về ngữ âm, đối lập về ý nghĩa, biểu hiện các khái niệm tương phản về logic, nhưng tương liên lẫn nhau. Việc xác định từ trái nghĩa cũng như từ đồng nghĩa của một từ sẽ giúp cho việc phân tích và sử dụng ngôn ngữ được chính xác hơn.

3.3.2. Ràng buộc ngữ nghĩa (semantic constraint)

Trong quá trình tạo câu, ngoài việc câu phải có đầy đủ các thành phần (đúng ngữ pháp) còn đòi hỏi các thành phần câu phải có mối liên kết, ràng buộc ngữ nghĩa lẫn nhau. Chỉ có xác lập được mối liên kết, ràng buộc ngữ nghĩa thì mới nhận ra được câu "*xe ăn cơm*" là không bình thường.

```
bắt V
   ...
   syntactics
        |
        +--category : V
        |
        +--subcategory : Vt
        |
        +--verb pattern : Sub+V+Obj+Obj
   semantics
        |
        +--logical constraint
        |         |
        |         +--category meaning : Action
        |         |
        |         +--synonym : buộc, ép
        +--semantic constraint
                  |
                  +--sub : Person
                  |
                  +--obj : LivingThing
                  |
                  +--obj : VP
   def : khiến phải làm việc gì, không cho phép làm khác đi.
   exa : bà bắt cháu đi ngủ ~ ông bắt trâu cày thông tầm.
```

Hình 5. Thông tin *Semantics* của "bắt" đòi hỏi hai bổ ngữ.

Do có vai trò quan trọng trong tiến trình phân tích ngôn ngữ nên các thông tin về *semantic constraint* và *logical constraint* thường được sử dụng để tạo ra các bộ luật phân tích cú pháp.

Ngoài các thông tin đã nêu, VCL còn đưa thêm 2 thông tin là *lời định nghĩa* (definition) và phần *ví dụ* (example) minh hoạ. Lời định nghĩa nêu lên ý nghĩa cơ bản của đơn vị từ vựng được khái quát từ những cảnh huống cụ thể trong hoạt động ngôn ngữ. Ví dụ là trường hợp vận dụng từ ngữ cụ thể được nêu ra để minh hoạ hoặc chứng minh cho lời định nghĩa. Hai thông tin này giúp cho người xây dựng từ điển VCL mô tả các thông tin liên quan khác được chính xác.

4. QUY TRÌNH XÂY DỰNG VCL

4.1. Tổ chức dữ liệu từ điển

Chúng tôi dựa vào quyển Từ điển tiếng Việt (2007) do Trung tâm Từ điển học phát hành để xây dựng nội dung cho VCL. Nói chung, trong quyển từ điển này, quan điểm về thu thập từ vựng, về chuẩn hoá chính tả, về chú thích từ loại, từ đồng âm, từ trái nghĩa là tương đối rõ ràng và thống nhất. Chúng tôi tách mỗi nghĩa của một đơn vị từ vựng được biểu diễn thành một mục từ (entry) trong VCL, khôngphân biệt là từ đồng âm hay từ đa nghĩa. Đồng thời, chúng tôi cũng

tách từ loại *kết từ* được nêu trong Từ điển tiếng Việt (2007) thành 2 loại *giới từ* và *liên từ*; tách danh từ chỉ số lượng thành *số từ*. Hiện tại, VCL chứa gần 42.000 mục từ. Toàn bộ dữ liệu từ điển VCL được tổ chức thành cơ sở dữ liệu, cho phép cập nhật, thay đổi khi cần thiết. Từ cơ sở dữ liệu này có thể dễ dàng biến đổi từ điển theo chuẩn XML.

4.2. Công cụ xây dựng VCL

Việc thiết kế một công cụ giúp cho quá trình xây dựng nội dung VCL là rất cần thiết. Công cụ cho phép tích hợp một số tiện ích như tạo mối quan hệ giữa 2 bộ nhãn từ loại, giữa 20 lớp ngữ nghĩa cơ sở với gần 100 tiểu loại của chúng trong cây phân loại ngữ nghĩa, v.v. Công cụ cũng cho phép tổ chức làm việc theo nhóm, làm việc theo từng vấn đề, do vậy công việc kiểm tra, đánh giá kết quả sẽ thuận lợi hơn.

4.3. Kho văn bản

Trong phân tích ngôn ngữ, một yêu cầu không thể thiếu đó là phải đặt đơn vị ngôn ngữ đang xét trong một tập hợp nói chung những đơn vị ngôn ngữ đứng trước và đứng sau nó. Tập hợp những đơn vị ngôn ngữ như vậy được gọi là ngữ cảnh. Như vậy, ngữ cảnh là một phương tiện để phân tích ngôn ngữ. Kho văn bản (corpus) được tổ chức là nguồn ngữ liệu hữu dụng phục vụ cho việc tìm ra ngữ cảnh của đơn vị ngôn ngữ.

Để giúp cho việc mô tả thông tin trong VCL, chúng tôi xây dựng một kho văn bản tiếng Việt, theo đó chúng tôi cũng thiết kế một công cụ dùng để tìm ngữ cảnh (Concordance).

5. KẾT LUẬN

Bài báo đã trình bày một cách tổng quan về việc xây dựng Từ điển tiếng Việt dùng cho máy tính. Qua đó đã đề xuất một mô hình cấu trúc và các bước cần thiết trong quá trình thiết kế, hoàn thành nội dung cho từ điển. Một cấu trúc đưa ra như vậy chắc chắn chưa thể đầy đủ cho các nhu cầu phân tích, miêu tả tiếng Việt. Tuy nhiên, với những kết quả ban đầu, chúng tôi hi vọng VCL sẽ được ứng dụng có hiệu quả ngay trong các đề tài về xử lí tiếng Việt.

Với mong muốn tạo ra một từ điển điện tử tiếng Việt tương thích với các từ điển khác, vấn đề cấu trúc của VCL sẽ được tiếp tục nghiên cứu, mở rộng trong tương lai. Chẳng hạn, bổ sung thông tin về từ (cụm từ) tương đương của tiếng nước ngoài (equivalent); thông tin về hình dạng (shape), kích cỡ (size) của các từ chỉ vật thể; thông tin về quan hệ giữa cái chỉnh thể và cái bộ phận (Whole-of), giữa cái bộ phận và chỉnh thể (Part-of), và những thông tin khác nếu thấy có nhu cầu ứng dụng trong các đề tài có liên quan đến nghiên cứu, xử lí tiếng Việt.

Lời cảm ơn: Việc xây dựng từ điển VCL được sự hỗ trợ kinh phí từ đề tài Nhà nước KC.01.01/06-10. Chúng tôi xin trân trọng cảm ơn sự giúp đỡ, tạo điều kiện từ phía Ban Chủ nhiệm Đề tài. Tập thể tác giả cũng xin chân thành cảm ơn các nhóm tham gia Đề tài đã góp nhiều ý kiến bổ ích trong quá trình thiết kế từ điển, cảm ơn các bạn đồng nghiệp ở Trung tâm từ điển học đã đóng góp nhiều công sức cho việc xây dựng từ điển.

TÀI LIỆU THAM KHẢO

Charoenporn T. (2004), *TCL' s Computational Lexicon*. Myanmar-Thai Co-Workshop on Myanmar Language Implementation MICT Park, Yangon Myanmar.

Hoàng Phê (2008), *Tuyển tập ngôn ngữ học*, Nhà xuất bản Đà Nẵng – Trung tâm Từ điển học.

ISO/TC 37/SC 4 N330 (Rev.13-2006, Rev.16-2008), *Language resource management - Lexical markup framework* (LMF).

Miller G., Backwith R., Fellbaum C., Gross D., Miller K. (1990), *Five papers on WordNet*, Technical report, Cognitive science laboratory, Princeton University.

Nguyễn Kim Thản (1997), *Nghiên cứu ngữ pháp tiếng Việt*, Nhà xuất bản Giáo dục.

Nguyễn Minh Thuyết, Nguyễn Văn Hiệp (2004), *Thành phần câu tiếng Việt*, Nhà xuất bản Giáo dục.

Nguyen T. M. H., Vu X. L., Romary L., Rossignol M. (2007), *A Lexicon for Vietnamese Language Processing*, LRE (Language Resources and Evaluation), Special Issue: Asian Language Resources.

Nguyen T. M. H. (2006), Outils et Ressources Linguistiques pour l'alignement de textes de textes multilingues français-vietnamiens, Thèse de doctorat en Informatique, Université Henri Poincaré - Nancy I, France.

Vũ Xuân Lương (2002), *Thiết lập giao diện biên soạn từ điển ngôn ngữ trên máy tính*, Tạp chí Ngôn ngữ, Số 7.