CẤU TRÚC (TỪ ĐIỂN) OALD8 ANH - VIỆT DƯỚI GÓC ĐỘ TỪ ĐIỂN HỌC TÍNH TOÁN

ĐINH ĐIỀN *

***Tóm tắt*:** Đến nay, đã có rất nhiều từ điển Anh-Việt với dung lượng từ, nội dung mục từ và cách xây dựng khác nhau với chất lượng cũng khác nhau. Chất lượng của từ điển phụ thuộc vào nhiều yếu tố, như: cấu trúc vĩ mô, cấu trúc vi mô, các phương tiện phi ngôn ngữ,... Trong bài báo này, chúng tôi sẽ phân tích các yếu tố nói trên cho phiên bản Anh-Việt chính thức đầu tiên của từ điển OALD8 (Oxford Advanced Learner's Dictionary, 8th edition) Anh - Anh dưới góc độ từ điển học tính toán. Dựa trên các kết quả phân tích này, chúng ta có thể rút ra được những điểm cần chú ý khi đánh giá hay xây dựng một từ điển Anh - Việt theo tiếp cận liên ngành này.

***Từ khóa*:** Từ điển Anh - Việt, từ điển học tính toán, cấu trúc vi mô, cấu trúc vĩ mô.

***Abstract*:** So far, many English-Vietnamese dictionaries are available with various capacities, contents of entries and ways of building. The quality of dictionaries depends on many factors, e.g.: the macro-strucure, the micro-structure, the non-linguistic means, etc. In this article, we will analyze the above-mentioned factors in the first official version of the English-Vietnamese dictionary OALD8 (Oxford Advanced Learner's Dictionary, 8th edition) from the computing lexicography perspective. Basing on these analyzed results, we will be able to draw remarkable issues in evaluating or building an English-Vietnamese dictionary.

***Keywords*:** English-Vietnamese dictionary, micro-structure, macro-structure.

1. Mở đầu

Trong việc học ngoại ngữ nói chung và tiếng Anh nói riêng, một trong những yếu tố quan trọng quyết định sự thành bại đó chính là từ điển. Việc nghiên cứu và biên soạn từ điển đã có truyền thống hàng ngàn năm qua, nhưng trong thời đại Công nghệ thông tin, thời đại của Cách mạng Công nghiệp 4.0 hiện nay, việc biên soạn từ điển đã thay đổi rất nhiều. Trước đây, để xây dựng từ điển, các nhà làm từ điển phải tự sưu tầm thủ công các mục từ, các cách dùng từ, tự tham khảo/biên soạn các định nghĩa, tự tìm kiếm các ví dụ và tự tổ chức sắp xếp thủ công hàng vạn mục từ,… Cách thức thủ công này vừa chậm, tốn nhiều công sức, lại vừa dễ sai sót và lại không kịp cập nhật hết những từ mới, những cách dùng mới của từ. Trong khi đó, với sự trợ giúp của máy tính và các ngành có liên quan máy tính như: ngôn ngữ học tính toán, ngôn ngữ học ngữ liệu, trí tuệ nhân tạo,…, phần lớn các công việc đó đều đã được tự động hóa. Theo cách tiếp cận mới này, từ điển vừa được xây dựng nhanh chóng, hầu như không sai sót mà lại cập nhật nhanh chóng những từ mới, những cách dùng mới của từ. Đây chính là cách tiếp cận của liên ngành giữa Từ điển học và Tin học được gọi là *Từ điển học Tính toán* (Computational Lexicography) [1] (viết tắt là

* *PGS.TS - Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Tp. Hồ Chí Minh; Email: ddien@fit.hcmus.edu.vn*

TĐHTT). Trong bài viết này, chúng tôi sẽ giới thiệu các kết quả phân tích về từ điển OALD8 ấn bản Anh - Việt [3] dưới góc độ TĐHTT và từ đó, chúng ta sẽ rút ra được các điểm cần chú ý khi đánh giá ưu khuyết điểm của một từ điển hiện nay, đồng thời rút ra được những kinh nghiệm quý báu trước khi bắt tay xây dựng một từ điển mới theo tiếp cận liên ngành này.

2. Tổng quan

Trên thị trường từ điển Anh - Việt hiện nay, có rất nhiều loại từ điển khác nhau, đa dạng về chủng loại, kích cỡ, cách tiếp cận,… của nhiều tác giả: Lê Bá Kông - Lê Bá Khanh, Lê Khả Kế, Đặng Chấn Liêu, Nguyễn Văn Khôn, Bùi Ý, Bùi Phụng,… cho đến hàng loạt các tác giả gần đây. Trong những quyển từ điển nổi tiếng trước đây, các nhà biên soạn từ điển đã xây dựng tiêu chí riêng để lựa chọn mục từ: những đơn vị nào được đưa vào từ điển, chọn tỉ lệ phù hợp giữa các từ cổ, từ cũ, từ thải loại, tiếng lóng, tiếng địa phương, từ chuyên ngành so với vốn từ toàn dân, từ phổ biến. Bên trong mỗi mục từ, nội dung giải thích sẽ được cân nhắc kỹ lưỡng những thông tin nào cần đưa vào, những thông tin nào không, cách giải thích như thế nào, ví dụ ra sao,… đều tùy thuộc vào quan điểm nhất quán của nhà làm từ điển ban đầu cho phù hợp với đối tượng mà họ muốn hướng đến. Do đó, trong từng từ điển khác nhau với đối tượng khác nhau sẽ có tiêu chí lựa chọn mục từ khác nhau, cách giải thích bên trong nội dung của mỗi mục từ cũng sẽ khác nhau.

Tuy nhiên, có một thực tế là các từ điển nổi tiếng của các học giả nói trên, đã được xuất bản cách đây trên dưới nửa thế kỷ không thể cập nhật các từ mới hay các cách dùng mới trong vài chục năm gần đây. Còn các từ điển Anh - Việt mới xuất bản gần đây, đa số được biên soạn, tổng hợp dựa trên các cuốn cũ trước đó hoặc dịch từ những cuốn từ điển Anh - Anh gần đây. Trong quá trình biên soạn hay biên dịch đó, nhằm tạo ra một từ điển mới có khối lượng từ nhiều hơn và quan trọng là nhằm tránh vấn đề vi phạm bản quyền (theo Công ước Bern), các tác giả mới phải chọn phương thức tổng hợp từ nhiều quyển từ điển trước đó và mỗi quyển được rút trích một phần hay cố tình chỉnh sửa cho khác đi các định nghĩa, các lời giải thích, các ví dụ,… Với cách làm này, họ có thể tạo ra các từ điển Anh - Việt mới đồ sộ hơn, nhiều từ hơn. Điều này vô hình chung gây ra sự không nhất quán, mất đi tính hệ thống của các từ điển cũ nổi tiếng trước đó mà họ đã dựa vào. Ngoài ra, với cách xây dựng thủ công, từ điển mới tuy đồ sộ nhưng lại bao gồm rất nhiều từ cũ, cổ, cách dùng không thông dụng mà lại thiếu từ mới hay cách dùng mới, đặc biệt là không có sự sắp xếp về tần suất sử dụng của từ. Chính điều này khiến cho người sử dụng sẽ khó tra cứu, khó liên hệ với những từ có liên quan với nhau, cách giải thích lòng vòng, khó hiểu, không nhất quán,… làm giảm lòng tin của người sử dụng từ điển mới đó.

3. Sơ lược về *Từ điển OALD8 Anh - Việt*

OALD8 (Oxford Advanced Learner's Dictionary, ấn bản lần thứ 8) là từ điển Anh - Anh nổi tiếng thế giới của Nhà xuất bản Đại học Oxford và là từ điển Anh - Anh bán chạy nhất thế giới hiện nay (hơn 35 triệu bản đã bán ra). Với hơn 100 năm kinh nghiệm biên soạn từ điển tiếng Anh, OUP đã biết tận dụng những thành quả mới nhất của chuyên ngành từ điển học tính toán trong việc thu thập mục từ rõ ràng, hợp lý, biên tập, tổ chức nội dung mục từ rất khoa học và ngày càng phong phú hóa phần phụ lục với những chủ điểm rất thiết thực.

OALD8 Anh - Anh là từ điển tường giải, dành cho người học tiếng Anh có trình độ cao, do đó, OUP đã xây dựng tiêu chí lựa chọn mục từ và cách giải thích trong nội dung mục từ phù hợp với đối tượng đó. Để xây dựng được danh sách mục từ hợp lý, các ví dụ phù hợp với từng cách dùng từ, OUP đã khai thác các

thành quả nghiên cứu mới nhất về TĐHTT và thực nghiệm trên hai kho ngữ liệu tiếng Anh nổi tiếng, đó là: BNC (British National Corpus) và OEC (Oxford English Corpus). Vì là từ điển tường giải, nên phần định nghĩa rất quan trọng. Để đảm bảo được *độ khó* (readability) của câu văn trong phần định nghĩa thấp hơn mức cho phép, OUP đã nghiên cứu xây dựng danh sách 3.000 từ cơ bản gồm những từ được sử dụng phổ biến nhất, dễ hiểu nhất dựa vào kết quả thống kê của ngôn ngữ học ngữ liệu từ những kho ngữ liệu thực tế nói trên. Tất cả các định nghĩa,

OALD8 Anh - Việt [3] (Oxford Advanced Learner's Dictionary 8th ed with Vietnamese Translation) là từ điển song ngữ Anh - Việt được biên dịch chính thức (có bản quyền từ OUP) từ *OALD8 Anh - Anh* gốc nói trên và được công bố vào đầu tháng 3.2015 [4]. Trong các phần dưới đây, chúng tôi sẽ phân tích chi tiết về cấu trúc vĩ mô, cấu trúc vi mô và phần phụ lục của từ điển mới này dưới góc độ của nghiên cứu liên ngành TĐHTT, đồng thời có sự so sánh với các từ điển Anh-Việt khác hiện nay để rút ra những bài học về làm từ điển.

3. Tổng quan về liên ngành *Từ điển học tính toán*

Đây là liên ngành giữa Từ điển học và Tin học nhằm sử dụng các công cụ, thành quả nghiên cứu bên Tin học trong lĩnh vực Ngôn ngữ học Tính toán (Computational Linguistics) để phục vụ cho việc nghiên cứu và thực hành Từ điển học [2]. TĐHTT không đơn thuần là sử dụng tin học để sắp xếp, trình bày, xuất bản từ điển trên môi trường điện tử mà quan trọng là phương pháp luận, cách tiếp cận, thiết kế, biên soạn, kiểm tra và lưu trữ đều mang tính định lượng, tự động hóa tối đa, khách quan, nhất quán, tính mở (open) và tính động (dynamic). Từ điển học có liên quan và thừa hưởng các kết quả nghiên cứu của nhiều chuyên ngành hay liên ngành khác của ngôn ngữ học, như: ngữ âm học, từ vựng học, hình thái học, ngữ nghĩa học, ngữ dụng học, phong cách học,… Vì vậy, TĐHTT cũng có liên quan và thừa hưởng các kết quả nghiên cứu của ngữ âm học tính toán, từ vựng học tính toán, hình thái học tính toán, ngữ nghĩa học tính toán, ngữ dụng học tính toán, phong cách học tính toán,… và tất cả các ngành hẹp này đều thuộc về Ngôn ngữ học Tính toán, một liên ngành giữa Ngôn ngữ học và Máy tính.

Chẳng hạn, trong việc xây dựng bảng từ, cách tiếp cận của từ điển học truyền thống là dựa trên lý thuyết về từ vựng học, họ đưa ra tiêu chí nhận diện từ và sử dụng tiêu chí đó để lựa chọn thủ công các đơn vị mà họ cho là đúng với tiêu chí đó, để đưa vào bảng từ. Việc lựa chọn thủ công từ một danh sách rất lớn các ứng viên dễ dẫn đến tính không nhất quán, mang tính chủ quan, cảm tính của người chọn. Mặt khác, danh sách các ứng viên dù rất lớn, nhưng vẫn khó mà đảm bảo bao quát hết các mục từ có thể có trong ngôn ngữ đó. Điều này đã xảy ra trong thực tế khi có nhiều từ điển tiếng Việt chúng ta (dù là uy tín) nhưng vẫn thiếu nhiều mục từ (thậm chí hàng trăm mục từ), trong đó có nhiều mục từ thông dụng, như: *truyền thông, tiếp thị, doanh gia, doanh nhân, tin tặc, kết xuất, in ấn, máy in, đơn ngữ, ngữ dụng,…*

Vì danh sách ứng viên đáp ứng đúng tư cách là từ thường vẫn còn rất lớn (có thể lên đến hàng trăm ngàn, có rất nhiều từ địa phương, từ cổ, từ chuyên ngành, từ vay mượn, từ mới,…) và chúng ta không thể đưa hết vào bảng từ, mà phải lựa chọn những mục từ ứng viên theo độ phổ biến từ trên xuống. Do đó, nếu lựa chọn thủ công, kết quả sẽ không chính xác vì theo cảm tính của người chọn mà cảm tính này lại phụ thuộc vào xuất thân, trình độ, vùng miền, chuyên môn,… của người chọn. Điều này dẫn đến có những mục từ xứng đáng hơn nhưng thiếu và ngược lại. Tương tự cho việc xem xét hết các cách dùng của một từ để người biên soạn từ điển có cách định nghĩa

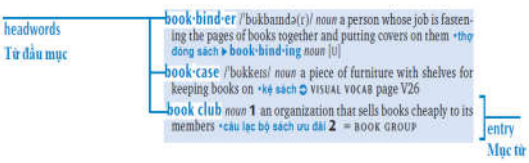
phù hợp cho từng nghĩa và kèm theo ví dụ minh họa cho cách dùng đó. Với từ điển học truyền thống, họ phải đọc thủ công từ rất nhiều tài liệu khác nhau, tác phẩm khác nhau ở nhiều lĩnh vực, phong cách khác nhau hoặc họ tham khảo từ những từ điển xuất bản trước đó (thậm chí trước vài chục năm) để ghi lại các câu ví dụ về cách dùng của một mục từ nào đó. Với cách này, vì sức con người có hạn, nên chỉ có thể thu thập thủ công tối đa lên đến vài trăm ngàn phiếu trong vòng hàng chục năm và trong đó có nhiều cách dùng cũ. Điều này sẽ dẫn đến thiếu rất nhiều các cách dùng khác ở những lĩnh vực khác hay cách dùng mới mà người biên soạn chưa kịp thu thập, cập nhật.

Trong khi đó, với TĐHTT [1], tất cả các bài toán trên đều được giải quyết tự động, định lượng và khách quan trên hàng tỷ phiếu điện tử. Do đó, việc quan sát cách dùng từ trong TĐHTT trở nên khoa học, chính xác và khách quan hơn. Để giải quyết các bài toán trên, TĐHTT sẽ xây dựng kho ngữ liệu cho từ điển bằng cách thu thập tự động (bằng công cụ tin học) từ hàng triệu website có liên quan đến lĩnh vực, thể loại của từ điển cần xây dựng. Kho ngữ liệu này có kích thước lên đến vài tỷ từ, chứa vài trăm triệu câu, rất phong phú, đa dạng, bao quát hầu hết các cách dùng từ trong những thể loại, vùng miền, lĩnh vực, niên đại, phong cách khác nhau. Từ kho văn bản thô này, chúng ta mới tiến hành xây dựng kho ngữ liệu theo những tiêu chí của ngôn ngữ học ngữ liệu bằng các công cụ tin học để chọn lựa văn bản theo từng thể loại, chủ đề, vùng miền, niên đại,… sao cho đảm bảo tính cân bằng (balance), tính đại diện (representative) của kho ngữ liệu sau này.

4. Cấu trúc vĩ mô của *OALD8 Anh - Việt*

Trong OALD8 Anh - Việt, cấu trúc vĩ mô của từ điển được bảo toàn nguyên trạng từ cấu trúc vĩ mô của từ điển gốc OALD8 Anh - Anh. Theo đó, cấu trúc vĩ mô này là danh sách các từ đầu mục (headword) và được sắp xếp theo trật tự chữ cái ABC của từ đầu mục. Hình thái của từ đầu mục là hình thức phổ biến nhất được thống kê từ kho ngữ liệu BNC, OEC. Tiêu chí lựa chọn từ đầu mục dựa trên mục đích sử dụng của người đọc (đối tượng sử dụng từ điển ở đây là người học tiếng Anh có trình độ cao). Do đó, tỉ lệ các từ hàn lâm sẽ cao hơn, nhưng vì đây là từ điển tổng quát, nên tỉ lệ các từ chuyên ngành sâu sẽ thấp, ngoại trừ những từ chuyên ngành đã được phổ biến rộng rãi. Tỉ lệ các từ cổ, tiếng lóng, từ dùng lâm thời, từ thông tục thấp. Tóm lại: tỉ lệ giữa các từ cổ, từ lóng, từ vay mượn, từ chuyên ngành sâu,… so với từ toàn dân, tổng quát có tỉ lệ hợp lý. Do đó, dù số lượng mục từ trong OALD8 thấp (chỉ khoảng 70.000 mục từ), nhưng trong mỗi mục từ, lại chứa nhiều từ phái sinh, nên khiến vốn từ lên đến 184.500 từ. Cách tổ chức bảng mục từ như thế là hợp lý, để người sử dụng dễ dàng tra cứu, liên hệ những từ có liên quan với nhau theo cách quan hệ: đồng nghĩa, trái nghĩa, tương tự, phái sinh, … (*xem Hình 1*).

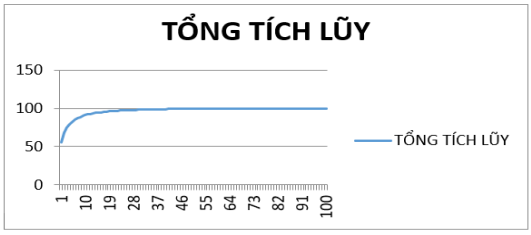
Ngoài các đơn vị mục từ ở cấp độ *từ* (word) là chính, trong từ điển này còn có các đơn vị trên từ như: *ngữ động từ* (phrasal verb), *thành ngữ* (idiom), *tục ngữ* (proverb), ... Tuy nhiên, khác với các từ điển tiếng Việt, trong OALD8 không có các đơn vị dưới từ (như: hình vị).



Hình 1. Cấu trúc bảng từ của OALD8 Anh - Việt

Với những mục từ thuộc lớp 3.000 từ cơ bản sẽ được đánh dấu bằng biểu tượng chìa khóa, để thể hiện đây là những từ chính, có tần suất sử dụng cao nhất (keyword). Với những

từ thuộc lớp từ hàn lâm cơ bản (để phục vụ trong việc đọc tin tức chuyên ngành), mục từ sẽ được đánh dấu bằng ký hiệu AW (academic word). Cũng có những từ thuộc đồng thời cả 2 lớp trên. Việc đánh dấu lớp 3.000 từ cơ bản này có ý nghĩa lớn đối với người sử dụng vì nó chính là lớp các từ có độ phổ biến hay tần suất sử dụng cao nhất và cũng là lớp các từ giúp làm giảm độ khó của văn bản mạnh nhất, có ảnh hưởng rõ nét nhất lên việc đọc của người sử dụng [5]. Đây chính là một trong những thành quả cụ thể mà TĐHTT đã mang lại cho việc biên soạn từ điển theo công nghệ tính toán bằng máy tính. Độ khó của văn bản tùy thuộc vào độ phổ biến của từ, kết cấu câu và tổ chức văn bản. Riêng trong trường hợp các câu định nghĩa trong từ điển, thì độ khó này không phụ thuộc và tổ chức văn bản (vì thường định nghĩa chỉ gồm một câu). Tần suất xuất hiện các âm tiết xuất hiện với tần suất khác nhau. Độ phổ biến của từ phụ thuộc vào tần suất xuất hiện của từ đó trong kho ngữ liệu được thu thập theo đúng các chuẩn về tính đại diện, tính cân bằng và kỹ thuật lấy mẫu. Kho ngữ liệu này phải đủ lớn và bao quát các lĩnh vực khác nhau mà lĩnh vực của từ điển tính đến [6]. Độ phổ biến này được chuẩn hóa bằng công thức f = –lg (n/N) với n là số lần xuất hiện của đơn vị ngôn ngữ đó (ví dụ ở đây là từ) và N là tổng số đơn vị đó xuất hiện trong toàn kho ngữ liệu. Ví dụ: trong kho ngữ liệu chứa 100 triệu từ, từ "và" xuất hiện 1 triệu lần thì f sẽ là –lg(10exp6/10exp8) = 2. Nếu con số này càng nhỏ (nhỏ nhất là 0), có nghĩa là từ đó xuất hiện càng nhiều và ngược.



Hình 2. Đồ thị tần suất tích lũy

Từ kết quả thống kê tần suất xuất hiện của các đơn vị ngôn ngữ ở các cấp độ khác nhau (ví dụ ở đây là cấp độ từ), nhóm tác giả [5] đã rút ra được đồ thị tổng tích lũy như hình 2 trên. Theo đó, 1% các từ phổ biến nhất (khoảng 340 từ so với vốn từ 34.000 từ của tiếng Việt) sẽ chiếm khoảng 55% số lượt từ xuất hiện trong kho ngữ liệu tiếng Việt Vcor gồm hơn 330 triệu từ [5]. Tương tự, 10% các từ phổ biến nhất (khoảng 3.400 từ) sẽ chiếm hơn 90% số lượt từ sử dụng trong VCor. Kết quả thống kê này có ý nghĩa to lớn trong việc dạy tiếng Việt. Điều này có nghĩa là người học ban đầu chỉ cần học khoảng 3.400 từ cơ bản nhất trong tiếng Việt là có khả năng hiểu được 90% văn bản/hội thoại trong tiếng Việt.

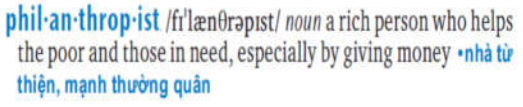
Với những từ đầu mục mà có biến thể (cách viết) khác, thì cách phổ biến hơn sẽ có nội dung giải thích đầy đủ, còn biến thể ít phổ biến hơn vẫn có mặt trong bảng từ, nhưng được liên kết đến nội dung đã giải thích trong mục từ chính (từ có cách viết phổ biến hơn) trên. Với những từ đầu mục thuộc dạng đồng tự - dị âm (cùng cách viết, khác cách đọc), thì ngay bên phải từ đầu mục đó sẽ có một chỉ số trên (superscript) để đánh số hai mục từ khác nhau nhưng có liên quan với nhau.

5. Cấu trúc vi mô của *OALD8 Anh - Việt*

Cấu trúc vi mô chính là nội dung bên trong mỗi mục từ (bắt đầu là từ đầu mục: headword). Đây là một hệ thống những thông tin có liên quan, như: phiên âm, chính tả, ngữ pháp, ngữ nghĩa, định nghĩa, phần dịch tiếng Việt, cách dùng. Ví dụ: từ nguyên, đồng nghĩa/phản nghĩa, phái sinh, tổ hợp cố định, khuôn mẫu,… của từ. Trong các trường thông tin kể trên, ngoại trừ các trường thông tin về phần dịch tiếng Việt, còn tất cả các nội dung còn lại đều được kế thừa nguyên bản của từ điển OALD8 gốc *OALD8 Anh - Anh*.

Về trường phiên âm (phonetics), OALD8 này chứa cả phiên âm giọng chuẩn Anh (Br.E)

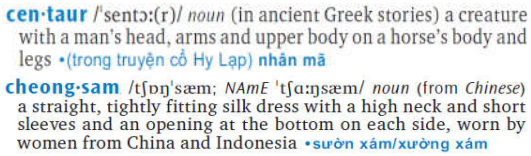
và giọng Bắc Mỹ (NAm.E). Với trường hợp có nhiều cách phát âm, cách phát âm phổ biến hơn sẽ được ghi trước. Về trường định nghĩa (definition), OALD8 này có điểm đặc biệt so với các từ điển khác là mọi câu định nghĩa đều được viết dựa trên 3.000 từ cơ bản đã nói ở trên nhằm tránh định nghĩa một từ bằng những từ khó hơn (như đã xảy ra với nhiều từ điển tiếng Việt). Một ví dụ cụ thể của việc áp dụng công nghệ trong việc biên soạn định nghĩa, giúp người sử dụng có thể đọc hiểu được định nghĩa một cách rõ ràng (*xem Hình 3*).


phil·an·throp·ist /fɪˈlænθrəpɪst/ *noun* a rich person who helps the poor and those in need, especially by giving money •nhà từ thiện, mạnh thường quân

Hình 3. Ví dụ về câu định nghĩa trong OALD8

Các nghĩa của từ được tổ chức rõ ràng theo từ loại, theo nghĩa gốc, theo lĩnh vực và có sự sắp xếp thứ tự theo mức độ thông dụng. Để lượng hóa được độ đo tính thông dụng người ta sử dụng công cụ thống kê kho ngữ liệu có gán nhãn thông tin ngôn ngữ như: hình thái từ, từ loại, ngữ nghĩa, ngữ dụng, phong cách, lĩnh vực,.... Việc sắp xếp thứ tự nghĩa của các từ đa nghĩa giúp ích rất nhiều cho người sử dụng từ điển để chọn nghĩa phù hợp với trình độ ngoại ngữ của mình (chẳng hạn theo thang đo 6 bậc của Khung tham chiếu đánh giá năng lực tiếng Anh của Châu Âu - CEFR).

Các ví dụ trong từ điển OALD8 này được rút ra từ những kho ngữ liệu thực tế nổi tiếng (BNC, OEC) và được lựa chọn sao cho phù hợp nhất với cách dùng từ cần minh họa. Trong nhiều từ điển Anh - Việt hiện nay, các ví dụ được lấy từ các từ điển khác nhau và có cải biên/tự chế (để tránh vấn đề đạo văn). Điều này có thể dẫn đến hiểu sai cách dùng từ (vì cùng một từ tiếng Anh, nhưng chỉ có thể dùng riêng cho con vật; có từ chỉ dùng cho phái nữ, …). Ví dụ: **bellow** *v.* (for a bull) to make a loud roar. **My boss got so mad he started **bellowing** at me.* Ngoài các ưu điểm (có sẵn từ từ điển gốc Anh - Anh), điểm đặc biệt của *OALD8 Anh - Việt* này chính là phần dịch tiếng Việt. Nhóm biên dịch đã tuân theo nguyên tắc ưu tiên tìm những từ ngữ tiếng Việt tương đương với mục từ tiếng Anh. Từ ngữ tiếng Việt tương đương phải phù hợp cả về mặt ngữ nghĩa, ngữ pháp và ngữ dụng (cách dùng). Ngoài ra, đây phải là từ toàn dân, từ đã được sử dụng rộng rãi chính thức (dựa trên từ điển tiếng Việt của Viện Ngôn ngữ học, ngữ liệu thực tế thu thập từ các báo điện tử chính thống, uy tín). Tuy nhiên, do sự khác biệt về loại hình văn hóa, loại hình ngôn ngữ, nên nhiều trường hợp rất khó và không thể tìm được từ/ngữ tương đương trong tiếng Việt, khi đó nhóm biên dịch sẽ dịch thêm lời giải nghĩa gốc của *OALD8 Anh - Anh* cho mục từ đó ra tiếng Việt để độc giả hiểu rõ thêm. Theo tôi, việc dịch tiếng Việt này là cần thiết, vì trong nhiều trường hợp, độc giả (do có trình độ tiếng Anh tốt) hoàn toàn có thể hiểu hết ý của câu định nghĩa gốc bằng tiếng Anh, nhưng khó tìm ra được ngay từ tương đương trong tiếng Việt. Ví dụ:


cen·taur /ˈsentɔ:(r)/ *noun* (in ancient Greek stories) a creature with a man's head, arms and upper body on a horse's body and legs •(trong truyện cổ Hy Lạp) nhân mã

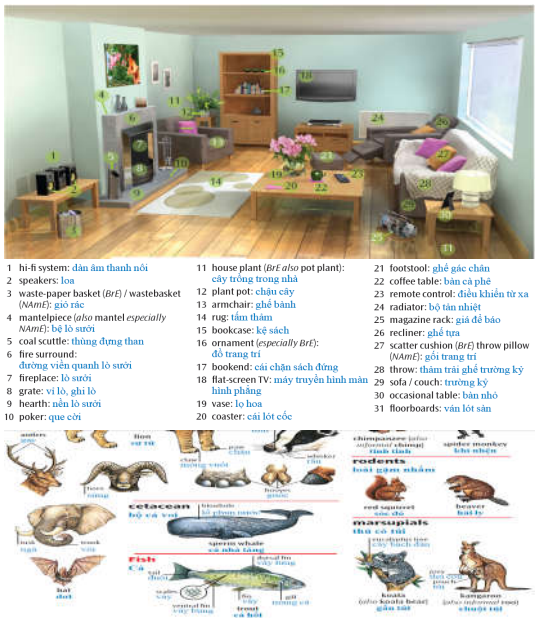
cheong·sam /tʃɒŋˈsæm; *NAmE* ˈtʃɑ:ŋsæm/ *noun* (from *Chinese*) a straight, tightly fitting silk dress with a high neck and short sleeves and an opening at the bottom on each side, worn by women from China and Indonesia •sườn xám/xường xám

Trong một số trường hợp, hình ảnh (một phương tiện phi ngôn ngữ) được sử dụng để minh họa thêm. Ngoài các mục từ truyền thống, trong từ điển này còn có nhiều hộp trợ giúp (boxes) để giúp người học hiểu hơn về cách dùng từ, phát triển từ vựng, như: Synonym/Antonym, Word Family, Collocation, British/American, Which Word, Grammar Point, Word Usage,…

6. Phần mục lục của *OALD8 Anh - Việt*

Điểm thiếu sót lớn nhất của các từ điển Anh - Việt trước đây chính là phần phụ lục

quá nghèo nàn (trong khi đây là những phương tiện phi ngôn ngữ trợ giúp người sử dụng từ điển nhiều nhất). Trong từ điển này, phụ lục bao gồm: hình vẽ, bảng biểu, tranh ảnh, bản đồ, hình ảnh kết cấu,… theo những chủ đề thiết thực trong cuộc sống (nhà cửa, thức ăn, phương tiện, thể thao, trò chơi, nhạc cụ,...) và có những chủ đề mới (biến đổi khí hậu,…) (*xem Hình 4*). Với sự trợ giúp của TĐHTT [7], OALD8 đã khai thác tối đa các phương tiện phi ngôn ngữ khi xây dựng phần phụ lục này và có sự liên kết 2 chiều giữa phần chính văn và phần phụ lục, giúp người sử dụng dễ dàng tra cứu hơn. Ngoài ra, trong OALD8 còn có phần "Oxford Writing Tutor" hướng dẫn cách viết như thể loại văn bản khác nhau: viết email, thư từ, báo cáo, luận,…



Hình 4. Các hình ảnh trong phần phụ lục

Một điều đáng chú ý ở từ điển này để các nhà làm từ điển chúng ta cần học hỏi, đó chính là việc ứng dụng công nghệ thông tin trong việc biên soạn, thiết kế, in ấn từ điển. Trong phần chính văn (cấu trúc vĩ mô, cấu trúc vi mô), toàn bộ nội dung được định dạng dưới tập tin kiểu có đánh dấu (mark-up) XML, DTD. Chính vì vậy, mà toàn bộ các trường thông tin khác nhau, các ký hiệu khác nhau (dù là nhỏ nhất) đều được quy định rõ ràng, chặt chẽ từ hình thức (font, size, style, indent, …) đến nội dung (những thông tin nào được ghi trong trường nào, thứ tự, quan hệ giữa các trường thông tin, ...). Chính nhờ các trường thông tin đã được cấu trúc hóa này, nên vi

7. Kết luận

Ngày nay, tiếp cận TĐHTT là xu thế chung của các thương hiệu từ điển lớn trên thế giới mà OALD8 là một ví dụ đã trình bày ở trên. Với tiếp cận này, máy tính trợ giúp đáng kể cho người xây dựng từ điển: từ khâu thu thập thẻ từ, xây dựng bảng từ, lựa chọn ví dụ cho cách dùng từ, biên soạn định nghĩa,... cho đến việc kiểm tra tính nhất quán, tính đầy đủ của các một hệ thống từ điển hàng trăm ngàn mục từ một cách tự động và chính xác.

Vì vậy, để xây dựng được một từ điển có chất lượng, theo tiếp cận TĐHTT, chúng ta phải xác định rõ đối tượng sử dụng, nhu cầu sử dụng, để từ đó đề ra tiêu chí thu thập kho ngữ liệu lớn phù hợp với lĩnh vực, thể loại, niên đại, vùng miền,... Sau đó, chúng ta phải xử lý, gán nhãn ngôn ngữ cho kho ngữ liệu đó. Từ kho ngữ liệu đã được gán nhãn thông tin ngôn ngữ, chúng ta khai thác để phục vụ cho hầu hết các công đoạn trong việc xây dựng từ điển: như xây dựng bảng từ với vốn từ phù hợp với trình độ và mục đích của người sử dụng; lựa chọn nghĩa, cách dùng từ phù hợp; câu định nghĩa có độ khó phù hợp; ... Tất cả điều này sẽ làm nên chất lượng của từ điển vì nó đáp ứng phù hợp nhất với nhu cầu của người sử dụng.

Ngoài ra, TĐHTT còn giúp chúng ta xây dựng các phiên bản điện tử của từ điển để người sử dụng có thể tra cứu nhanh chóng mọi lúc mọi nơi chứ không còn phụ thuộc với quyển từ điển giấy như trước đây.

⇒ *Xem tiếp trang 20*