

NGÔN NGỮ HỌC VÀ VIỆT NGỮ HỌC

TRẮC ĐẶC VỐN TỪ CỦA TRẺ EM BẰNG PHƯƠNG PHÁP NGÔN NGỮ HỌC NGỮ LIỆU PHẠM HIỀN* - PHẠM TIẾN DŨNG**

TÓM TẮT: Nghiên cứu xây dựng kho ngữ liệu cho trẻ em là việc làm cần thiết để làm cơ sở cho ngành ngôn ngữ cũng như trong lĩnh vực y học để xây dựng các bảng từ thử, câu thử đánh giá khả năng nghe hiểu lời nói của trẻ. Bảng phương pháp ngôn ngữ học ngữ liệu đã xây dựng được một kho ngữ liệu có kích thước 176.153 từ ngữ từ nguồn dữ liệu 232 truyện, sách dành cho trẻ em. Trắc đặc tương quan với tần số của một số kho ngữ liệu khác, chúng tôi tính toán số lượng từ ước tính cho từng lứa tuổi (từ 1 đến 6 tuổi). Đồng thời, bài báo trích rút ra 1.000 từ có tần suất xuất hiện cao nhất trong kho ngữ liệu cũng được phân tích để làm cơ sở ứng dụng cho các nghiên cứu sau này.

TỪ KHOA: kho ngữ liệu; ngôn ngữ trẻ em; tần số từ; độ phân tán từ vựng; tiếng Việt.

NHẬN BÀI: 22/8/2022.

BIÊN TẬP-CHỈNH SỬA-DUYỆT ĐĂNG: 5/10/2022

1. Giới thiệu

Trong nghiên cứu ngôn ngữ nói chung và nghiên cứu ngôn ngữ trẻ em nói riêng, từ lâu kho ngữ liệu đã được sử dụng như một nguồn tài nguyên cần thiết. Những dữ liệu sử dụng cho nghiên cứu ngôn ngữ trẻ em được lấy mẫu một cách hệ thống. Trước hết, có thể kể đến nguồn nhật ký của bố mẹ trẻ, sau đó là các đoạn băng ghi âm hoặc ghi hình. Bên cạnh những nghiên cứu sử dụng các kỹ thuật ngôn ngữ học thực nghiệm và ngôn ngữ học tâm lý, các kho ngữ liệu luôn là một cơ sở vững chắc đưa ra nhiều kết luận, đặc biệt là trong việc nghiên cứu các hiện tượng ngôn ngữ mới ở trẻ.

Trong nghiên cứu ngôn ngữ nói chung và ngôn ngữ trẻ em nói riêng, nhà nghiên cứu luôn cần ngữ liệu và càng nhiều ngữ liệu càng tốt. Kho ngữ liệu là tập hợp những tài liệu ngôn ngữ dưới dạng nói, viết thể hiện qua giao tiếp hàng ngày, báo viết, báo nói, báo hình, truyện, thơ ca, văn bản các loại. Việc nghiên cứu các kho ngữ liệu sẽ tính toán tần suất xuất hiện các từ theo loại từ khác nhau, độ phân tán của các từ qua đó có thể tìm ra các từ thông dụng được sử dụng nhiều nhằm mục đích xác định vốn từ cơ bản để xây dựng giáo trình, từ điển trong ứng dụng giảng dạy tiếng Việt. Trong lĩnh vực ngôn ngữ học nói chung và ngôn ngữ bệnh học nói riêng, việc nghiên cứu kho ngữ liệu là cơ sở để tìm ra các từ phổ thông, quen thuộc, có tần suất xuất hiện cao phù hợp với lứa tuổi của người nghe để đánh giá khả năng nghe hiểu lời nói trong chuyên ngành thính học. Chúng tôi tiến hành nghiên cứu với mục tiêu ứng dụng ngôn ngữ học ngữ liệu để xây dựng kho ngữ liệu các truyện, sách dành cho trẻ em nhằm mục đích trắc đặc vốn từ vựng của trẻ em.

Theo “Từ điển tiếng Việt” của Hoàng Phê, *trắc đặc* có nghĩa là “khảo sát, đo đạc, vẽ chi tiết một vùng đất nào đó” (tr.1026). Ở trong ngôn ngữ học ngữ liệu, cụ thể trong bài viết này, *trắc đặc* được hiểu là điều tra và đo lường ngữ liệu thu thập được để phục vụ mục đích nghiên cứu xác định trước đó. Trong bài báo này, chúng tôi sử dụng thuật ngữ “trắc đặc” với nội hàm nghiên cứu như trên.

2. Mục đích nghiên cứu

Từ việc xây dựng kho ngữ liệu ngôn ngữ trẻ em (chủ yếu là từ các tác phẩm văn học ở trong nước), bằng phương pháp của ngôn ngữ học ngữ liệu, chúng tôi đã tiến hành trắc đặc vốn từ của trẻ em, cụ thể là đưa ra Danh sách 1.000 từ ngữ có tần số xuất hiện nhiều nhất, Danh sách từ ngữ cốt lõi và dán nhãn từ loại cho chúng. Dựa trên kết quả trắc đặc đó, chúng ta có thể hình dung ra được một bức tranh rõ nét nhất về ngôn ngữ trẻ em. Kết quả nghiên cứu có thể được áp dụng làm cơ sở để nghiên cứu việc xây dựng danh sách từ phổ thông, từ cốt lõi trong lĩnh vực ngôn ngữ cũng như xây dựng danh sách từ thử ứng dụng trong y học để đánh giá khả năng nghe hiểu lời nói của cơ quan thính giác của trẻ em.

Để thực hiện được mục đích trên, trong bài báo này, chúng tôi trình bày vai trò của kho ngữ liệu trong nghiên cứu ngôn ngữ trẻ em. Quá trình xây dựng kho ngữ liệu truyện dành đọc cho thiếu nhi, bao gồm các truyện kể hằng đêm và các truyện và thơ đọc cho trẻ được thu thập làm tư liệu. Từ kho ngữ liệu đó, chúng tôi đã đưa ra một số phân tích bước đầu. Với dữ liệu vừa thu thập được ở trên,

* TS; Viện Ngôn ngữ học; Email: phamhieniol@gmail.com

** Bệnh viện Tai Mũi Họng Trung ương; Email: dungor176@gmail.com

chúng tôi coi đây là ngữ liệu đầu vào, tiếp đó, chúng tôi còn thu thập các lời nói trực tiếp của cha mẹ và người trông trẻ trong quá trình tiếp thụ tiếng mẹ đẻ của trẻ.

Sở dĩ chúng tôi tiến hành xây dựng kho ngữ liệu truyện thiếu nhi vì trước khi trẻ đi học ở trường, chúng đã tham gia vào nhiều hoạt động ngôn ngữ ban đầu như nghe đọc truyện và nghe kể chuyện. Trong những lần đọc to các truyện, trẻ em tham gia vào việc nghe âm thanh các từ các câu và nhìn vào tranh. Việc đọc to các cuốn sách được lựa chọn sẽ mang đến thế giới ngôn ngữ và học tập cho trẻ em, đồng thời tạo cơ hội thúc đẩy mối quan hệ khăng khít giữa người lớn và trẻ em. Trẻ mới biết đi thích đọc sách vì sự gần gũi về thể chất của chúng với người lớn và những cuộc trao đổi ngôn ngữ vui tươi diễn ra. Đọc sách với người lớn giúp họ cảm thấy an toàn và được yêu thương, và đó có thể là nền tảng của việc xây dựng các mối quan hệ tin cậy. Mặc dù đọc to cho trẻ nhỏ là một phần quan trọng của sự phát triển khả năng đọc viết, nhưng sẽ không đủ nếu không có sự tham gia của một người lớn phản ứng với trẻ và người mời trẻ tích cực tham gia sự kiện này cho dù trẻ còn nhỏ đến đâu. Chính vì tầm quan trọng cũng như ảnh hưởng to lớn của các sách đọc cho trẻ em tới quá trình thụ đắc ngôn ngữ của trẻ, cho nên, chúng tôi xây dựng kho ngữ liệu truyện trẻ em phục vụ cho việc nghiên cứu vốn từ vựng trẻ em nhìn từ cách tiếp cận ngôn ngữ học ngữ liệu.

Từ mục đích chung là có một kho ngữ liệu đủ lớn và cập nhật để khám phá sự đa dạng trong văn bản của trẻ em, bằng cách sử dụng tiềm năng ngày càng tăng của công nghệ siêu phương tiện, chúng tôi đã đi đến các mục tiêu cụ thể trong việc xây dựng kho ngữ liệu là: tạo ra một phiên bản điện tử của văn bản của các dự án dành cho trẻ em, để chúng có thể được khám phá bằng phần mềm phân tích văn bản, chẳng hạn như các chương trình phù hợp như một điều kiện tiên quyết để phát triển một kế hoạch nhất quán và minh bạch để phiên âm các văn bản của trẻ em, vì hầu hết chúng đều được viết tay. Từ kho ngữ liệu đó, chúng tôi tiến hành trích đặc các thông số về từ vựng cũng như ngữ pháp để cho thấy một bức tranh toàn diện nhất về ngôn ngữ của trẻ em nói tiếng Việt.

3. Phương pháp

Bài viết này sử dụng phương pháp ngôn ngữ học ngữ liệu để tiến hành nghiên cứu. Thuật ngữ “ngữ liệu” ở đây được dịch từ thuật ngữ tiếng Anh “corpus” (danh từ số nhiều là “corpora”) và trong tiếng Hi Lạp có nghĩa là “thân thể” (body). Nghĩa của từ “corpus” được hiểu là “phần thân của văn bản” và là tập hợp của nhiều văn bản. Chính vì vậy, từ “corpus” được dịch là “kho ngữ liệu”, hoặc cũng có các giả dịch là “khối liệu”. Ngữ liệu là những “dữ liệu, cứ liệu của ngôn ngữ”, tức là những chứng cứ thực tế sử dụng ngôn ngữ. Những chứng cứ sử dụng ngôn ngữ này có thể là của ngôn ngữ nói (spoken language) hoặc ngôn ngữ viết (written language). Ngữ liệu của ngôn ngữ viết thường được hiểu là tập hợp các văn bản. Hiện nay, các nhà ngôn ngữ học - ngữ liệu chủ yếu xét đến ngữ liệu dạng điện tử, tức là văn bản ở dạng máy đọc được (machine-readable texts). Ngữ liệu có thể tồn tại ở 2 dạng: dạng chỉ gồm các ngữ liệu thô thu được, không có chú thích (unannotated corpus) và dạng có chú thích (annotated corpus) thêm thông tin về ngôn ngữ cho các đơn vị ngôn ngữ trong ngữ liệu thô đó. Dạng thứ nhất thường là kết quả của việc thu thập văn bản dạng thô; còn dạng thứ hai là dạng văn bản thô đã được gán thêm thông tin về ngôn ngữ nhằm cho mục đích khai thác thông tin được hiệu quả hơn. Dưới đây chúng tôi trình bày các khâu tiến hành thu thập và xử lý dữ liệu.

3.1. Thu thập tư liệu

Chúng tôi lựa chọn ngữ liệu đầu vào từ các sách truyện dành cho thiếu nhi ở các độ tuổi. Các sách truyện này là những cuốn phổ biến và thường được lưu hành trong một số trường mẫu giáo cũng như gia đình của trẻ. Như vậy, có nghĩa là lượng từ vựng trong các sách truyện trên có vai trò trong việc trẻ đọc hoặc nghe truyện. Vốn từ vựng đó có vai trò lớn trong việc tiếp nhận và thụ đắc ngôn ngữ của trẻ. Vì những lý do trên, chúng tôi xây dựng tiêu chí chọn lựa các sách truyện như sau:

- + Về mặt thể loại: Chúng tôi chọn lựa thể loại truyện tranh, truyện dân gian, sách tiếng Việt lớp 1 tập 1 và tập 2 theo chương trình cũ thống nhất có 1 loại sách giáo khoa.
- + Về mặt uy tín của nơi xuất bản sách: Chúng tôi chọn lựa các sách truyện được các nhà xuất bản uy tín phát hành như các nhà xuất bản sau: Nhà xuất bản Giáo dục, Nhà xuất bản Kim đồng, Nhà xuất bản Trẻ.
- + Về thời gian xuất bản: Để có thể có độ cập nhật từ vựng tốt nhất, chúng tôi chọn lựa các sách truyện được xuất bản trong giai đoạn 2015-2020.

+ Về lứa tuổi của trẻ: Chúng tôi lựa chọn sách truyện dành cho trẻ em trong các độ tuổi từ 0-6 tuổi và các truyện cho trẻ có độ tuổi lớn hơn nhưng vẫn phù hợp với trẻ 6 tuổi theo khuyến cáo về độ tuổi phù hợp của các nhà xuất bản.

+ Về dạng thức tồn tại của văn bản đầu vào: Chúng tôi lựa chọn thể loại bản cứng không dùng bản điện tử do nó liên quan bản quyền và không kiểm soát được nội dung của bản điện tử.

+ Về độ phổ biến và dễ tiếp cận của các sách truyện: Chúng tôi chọn mua trực tiếp tại các cửa hàng giới thiệu sách của các nhà xuất bản, sau đó mới mua online để giúp gần với thực tế và khả năng tiếp cận truyện của trẻ.

+ Về tiêu chí tính dễ hiểu và thuần Việt của tác phẩm: Chúng tôi không chọn lựa các truyện dài tập theo chương, truyện hải, và không dùng các nhà xuất bản nước ngoài.

3.2. Xử lý tư liệu

Các văn bản sách truyện sau đó được chúng tôi tiến hành xử lý để có thể đạt tiêu chuẩn văn bản đầu vào cho kho ngữ liệu. Cụ thể, chúng tôi đã tiến hành các bước sau đây:

+ Dùng phần mềm chuyên dụng scan lại các tài liệu, chuyển thể sang dạng thuần văn bản sau đó đọc lại, rà soát các lỗi trước khi xử lý;

+ Làm sạch, chuẩn hoá và kiểm tra lỗi dữ liệu

+ Phân đoạn và gán nhãn ngôn ngữ cho dữ liệu;

+ Dùng phần mềm Word Smith 8.0 và một số chương trình xử lý và tính toán chuyên dụng;

+ Có so sánh và đối chiếu với từ điển tiếng Việt của tác giả Hoàng Phê, danh sách từ bao quát tiếng Việt;

+ Tính toán tần số và độ phân tán từ vựng;

+ Tính tần số từ vựng theo từ loại qua số lần xuất hiện của các từ đó trong kho ngữ liệu;

+ Gán nhãn từ loại cho kho ngữ liệu (gồm danh từ, động từ, tính từ, trạng từ, liên từ, sở từ, giới từ, đại từ, thán từ...).

Các tệp được kiểm tra và chuẩn hóa theo cùng một loại mã phông và định dạng thống nhất. Tất cả các tệp văn bản đã được chuyển đổi thành dạng Unicode UTF-8 để cho tương thích với các chương trình xử lý kho ngữ liệu, ví dụ: các chương trình tách từ và chương trình gán nhãn từ loại. Việc chuẩn hóa đã được áp dụng cho một số lỗi chính tả và lỗi bỏ dấu thanh không chính xác, ví dụ, *hỏa, lòe* và *tủy* được đổi thành *hoả, loè* và *tủy* tương ứng. Các biểu thức chính quy (regular expressions) đã được sử dụng để tìm và sửa các loại lỗi này. Như đã nói ở trên, tiếng Việt là một ngôn ngữ đơn lập phân tích tính điển hình do đó tiếng Việt có nhiều từ ghép. Chúng tôi đã sử dụng chương trình vnTokenizer (Lê et al. 2008) để phân đoạn các từ trong kho ngữ liệu. Chúng tôi cũng đã sử dụng chương trình vnTagger (Lê et al. 2010) để gán nhãn từ loại cho kho ngữ liệu. Vì các từ ghép này bản thân chúng hầu như luôn bao chứa các từ khác (ví dụ như trong tiếng Anh có từ *shower curtain* ‘rèm tắm’), tất cả các từ ghép đều được kết nối bằng dấu gạch dưới (_). Kết quả của hai quá trình này được lưu trữ riêng biệt thành ba phiên bản của kho ngữ liệu, đó là: phiên bản văn bản thường chưa được đánh dấu, phiên bản đã tách từ và phiên bản được gán nhãn cú pháp. Chúng tôi cung cấp thông tin về các chức năng ngữ pháp khác nhau của các từ bên cạnh tần số của chính từ đó. Kết quả là chúng tôi có được một kho ngữ liệu có thể sử dụng trong việc phân tích và tính toán các con số thống kê một cách chính xác và nhất quán bằng các chương trình khai thác kho ngữ liệu.

Từ quá trình thu thập ngữ liệu và xử lý tư liệu ở trên chúng tôi đã trắc đạc về mặt âm, từ vựng, ngữ pháp để đưa ra được bức tranh đặc trưng nhất về ngôn ngữ trẻ em dựa trên kho ngữ liệu ngôn ngữ trẻ em mà chúng tôi hiện đang có như đưa ra Danh sách 1.000 từ ngữ có tần số xuất hiện nhiều nhất, Bảng từ cốt lõi và Dán nhãn từ loại.

4. Kết quả

Sau khi đã văn bản hoá và chuẩn hoá cho các văn bản sách truyện từ sách giấy xuất bản, chúng tôi thu được 232 truyện, được lưu giữ thành các tệp văn bản điện tử thuần túy (plain text). Sau đó, chúng tôi tiến hành tính toán tần số, thứ hạng từ, lớp tần số, và một vài thông số khác. Các thông số đó phục vụ cho mục đích xác định các từ cốt lõi của trẻ em và 1.000 từ đánh giá âm trẻ em. Các thông số chúng tôi đã tính toán gồm tần số từ, độ phân tán từ vựng, thứ hạng từ, lớp tần số từ như có

thể thấy trong Bảng 2. Danh sách các truyện được chúng tôi sử dụng để xây dựng kho ngữ liệu có thể được truy cập tại địa chỉ sau:

<https://www.dropbox.com/s/so0ljx4bn1lorfh/Danh%20s%C3%A1ch%20truy%E1%BB%87n%20t%C3%ADnh%20t%E1%BA%A7n%20su%E1%BA%A5t%20t%E1%BB%AB.xlsx?dl=0>

Công thức tính Lớp tần số (Frequency Class - FC) như sau:

$$FC(w) = \log_2 \frac{freq_{max}}{freq(w)}$$

Lớp tần số được làm tròn lên số nguyên gần nhất. Lớp tần số $FC(w)$ của một từ w được tính toán theo mối tương quan giữa tần số của bản thân từ đó $freq(w)$ với tần số của từ có tần số cao nhất $freq_{max}$. Chẳng hạn, trong kho ngữ liệu đang xét, từ *con* là từ có tần số xuất hiện lớn nhất. Do đó, tần số của nó được sử dụng như là tần số tham chiếu $freq_{max}$. Năm từ có tần số cao nhất tiếp theo, *một*, *không*, *có*, *người* cũng nằm trong lớp tần số 0. Hai mươi một từ tiếp theo thuộc về lớp tần số 1. Số này tương ứng với 50% giảm xuống của tần số.

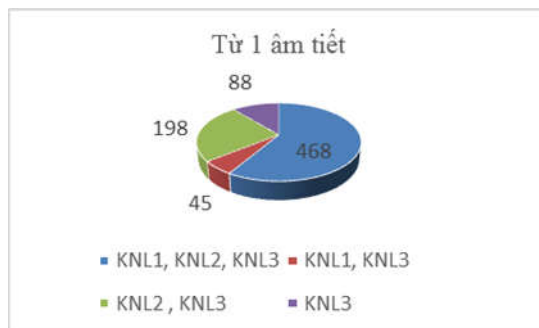
Kho ngữ liệu có kích thước 176.153 từ ngữ với 9.573 lượt từ được sắp xếp theo tần suất xuất hiện từ cao xuống thấp. Thống kê chỉ ra rằng có 3.191 từ 1 âm tiết còn lại là 6.382 từ ngữ từ 2 âm tiết trở lên. Tuy nhiên trong 1.000 từ có tần suất xuất hiện nhiều nhất thì lại có tới 799 từ 1 âm tiết và 201 từ ngữ 2 âm tiết.

Phân tích từ loại chỉ ra rằng danh từ gồm danh từ chung và danh từ riêng có số lượng 3.891 từ, chiếm 40,6%, động từ có 3.293 từ chiếm 34,3%, tính từ có 1758 từ chiếm 18,3%, trạng từ có 159 từ chiếm 1,6% còn lại là các từ loại khác.

Bảng 1. So sánh 1.000 từ ngữ có tần số xuất hiện cao nhất

Thông số \ Tác giả	Nguyễn Đức Dân	Phạm Giang	Phạm Hiền & Phạm Tiến Dũng
Từ 1 âm tiết	701	1000	799
Từ 2 âm tiết	299	0	201
Kích thước kho ngữ liệu	524500	207388	176153

Trong 799 từ 1 âm tiết trong nghiên cứu của chúng tôi có 468 từ trùng lặp khi so sánh đồng thời với kết quả nghiên cứu của tác giả Nguyễn Đức Dân (1981), Phạm Giang (2008), 513 từ trùng nếu chỉ so sánh với tác giả Nguyễn Đức Dân và 666 từ trùng nếu chỉ so sánh với nghiên cứu của tác giả Phạm Giang như có thể thấy trong Bảng 1 và Hình 2 dưới đây. Công trình của Nguyễn Đức Dân dựa trên kho ngữ liệu có nguồn từ các tài liệu gồm các bài báo, bài thơ, văn học cho trẻ em và một số tác phẩm của Hồ Chí Minh; trong khi đó, công trình của Phạm Giang dựa trên kho ngữ liệu có nguồn từ tài liệu từ các bài báo và văn học cho trẻ em. Như vậy có thể thấy rằng việc so sánh tần số từ kho ngữ liệu trong nghiên cứu của chúng tôi với hai kho ngữ liệu trên là có ý nghĩa để thấy sự khác biệt. Một điểm đáng chú ý là công trình của Phạm Giang không phân xuất từ ghép cho nên tần số tính toán trong công trình này là tần số âm tiết hoặc từ đơn, không có tần số từ ghép, như có thể thấy trong Bảng 1.



Hình 1. Phân tích sự trùng lặp từ 1 âm tiết (trong đó: KNL1: Nguyễn Đức Dân (1981); KNL2: Phạm Giang (2008); KNL3: Phạm Hiền & Phạm Tiến Dũng (nghiên cứu hiện tại))

Dưới đây là một số tiêu chí xác định từ vựng cốt lõi cho trẻ em theo từng độ tuổi:

1. Từ thường được sử dụng nhất trong ngôn ngữ trẻ em nói chung;
2. Từ thường được sử dụng nhất trong một phương tiện nào đó (ví dụ, trong ngôn ngữ nói người lớn hay dùng với trẻ em, hoặc trong ngôn ngữ viết trong cách sách và truyện dành cho trẻ em);
3. Từ hữu ích cho định nghĩa từ điển;
4. Từ phổ thông nhất, không mang tính không đánh dấu, hoặc là trọng yếu của ngôn ngữ;
5. Từ cơ bản về mặt nhận thức và nổi trội nhất;
6. Từ có cách sử dụng rộng rãi nhất trên nhiều thể loại với nghĩa phổ quát nhất của chúng;
7. Từ thường xuyên nhất cho một nhóm nhân khẩu học cụ thể.

Sau khi phân tích dữ liệu và rút ra được danh sách từ vựng theo tần số xuất hiện của từ trong kho ngữ liệu sách truyện trẻ em, chúng tôi kết hợp với các tiêu chí trên để chọn lựa ra một danh sách các từ cốt lõi của ngôn ngữ trẻ em.

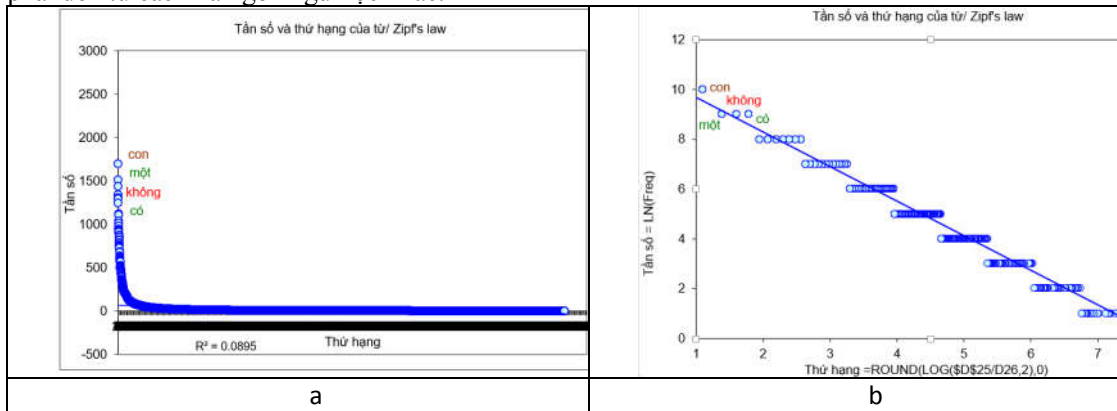
Bảng 2. Danh sách từ ước tính theo độ tuổi (trẻ 12 tháng tuổi, các từ bôi đậm là các từ cốt lõi)

Số thứ tự	Từ ngữ	Từ loại	Tần số	Tỉ lệ %	Tần số xuất hiện trong số tài liệu	Tỉ lệ %	Lớp tần số	Tuổi tương đối thụ đắc
1	CON	Danh từ chung	2,394	1.359	176	75.86	0	12 tháng
2	MỘT	Số từ	2,307	1.310	206	88.79	0	12 tháng
3	KHÔNG	Trạng từ	2,174	1.234	220	94.83	0	12 tháng
4	CÓ	Động từ	1,815	1.030	212	91.38	0	12 tháng
5	VÀ	Liên từ đẳng kết	1,804	1.024	214	92.24	0	12 tháng
6	NGƯỜI	Danh từ chung	1,695	0.962	177	76.29	0	12 tháng
7	ĐƯỢC	Trạng từ	1,508	0.856	182	78.45	1	12 tháng
8	ĐI	Động từ	1,432	0.813	211	90.95	1	12 tháng
9	CỦA	Giới từ	1,338	0.760	194	83.62	1	12 tháng
10	LẠI	Trạng từ	1,304	0.740	200	86.21	1	12 tháng
11	CHO	Động từ	1,294	0.735	207	89.22	1	12 tháng
12	LÀ	Động từ	1,241	0.705	201	86.64	1	12 tháng
13	BẠN	Danh từ chung	1,119	0.635	119	51.29	1	12 tháng
14	MẸ	Danh từ chung	1,107	0.628	132	56.90	1	12 tháng

Trong số các thông số tính toán kho ngữ liệu, thứ hạng từ là một thông số phản ánh định luật Zipf. Định luật Zipf mô tả tần suất của một từ trong ngôn ngữ tự nhiên, phụ thuộc vào thứ hạng của nó trong bảng tần số. Vì vậy, từ có tần số cao nhất sẽ xuất hiện cao gấp hai lần từ có lần số lớn thứ hai, và từ có tần số lớn thứ 2 cao gấp đôi từ có tần số lớn thứ tư, và như vậy cho đến khi từ ít thường xuyên nhất (xem Hình 3). Luật này được đặt tên theo nhà ngôn ngữ học người Mỹ George Kingsley Zipf (1902-1950), người đầu tiên đã cố gắng giải thích nó vào khoảng năm 1935. Zipf đã quan sát thấy rằng tần suất xuất hiện của các từ không được phân phối đồng đều cũng như bình thường, mà

thay vào đó tỉ lệ nghịch với thứ hạng tần số của chúng. Nghĩa là, sử dụng tham số phụ thuộc văn bản C , tần số của từ i tương ứng với việc phân chia C theo vị trí xếp hạng của i : $frequency_i = C / rank_i$

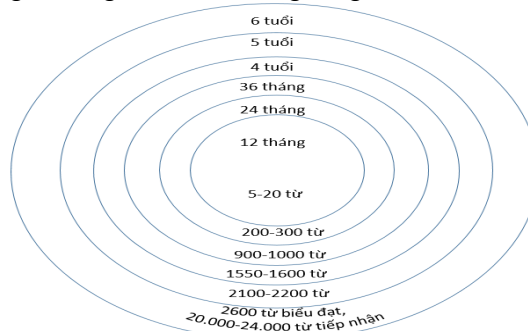
Hình 3 dưới đây thể hiện Phân bố Zipfian của tần số (trục dọc) và thứ hạng trong bảng tần số (trục ngang) của hàng trăm từ đầu tiên của kho ngữ liệu đang xem xét. Đường kẻ được dự đoán bởi định luật Zipf, và các dấu chấm tròn mô tả tần số từ thực tế trong kho ngữ liệu. Nhà ngôn ngữ học Lestrade của Trường Đại học Nijmegen, Hà Lan cho rằng “Tôi nghĩ có thể an toàn khi nói rằng định luật Zipf là bí ẩn lớn nhất trong ngôn ngữ học tính toán. Bất chấp nhiều thập kỉ đưa ra giả thuyết, nguồn gốc của nó vẫn khó nắm bắt”. Lestrade (2017) cho thấy rằng định luật Zipf có thể được giải thích bằng sự tương tác giữa cấu trúc của câu (cú pháp) và ý nghĩa của các từ (ngữ nghĩa) trong một văn bản. Sử dụng các mô phỏng máy tính, ông đã có thể chỉ ra rằng cả cú pháp hoặc ngữ nghĩa đều không đủ để tự tạo ra một bản phân phối Zipfian, nhưng cú pháp và ngữ nghĩa đó 'cân' lẫn nhau cho điều đó. Không chỉ các dự đoán dựa trên mô hình mới của Lestrades hoàn toàn phù hợp với các hiện tượng được tìm thấy trong ngôn ngữ tự nhiên, lí thuyết của ông còn phù hợp với hầu hết mọi ngôn ngữ trên thế giới, không chỉ đối với tiếng Anh hay tiếng Hà Lan. Lestrade cho biết ông vô cùng vui mừng với phát hiện này, và tôi bị thuyết phục về lí thuyết của mình. Tuy nhiên, sự xác nhận của nó phải đến từ các nhà ngôn ngữ học khác.



Hình 2. Biểu đồ mối quan hệ giữa tần số và thứ hạng từ. Hình (a) thể hiện giá trị thu được;

Hình (b) thể hiện giá trị tần số theo logarit và giá trị thứ hạng theo công thức

Sau khi áp dụng các tiêu chí nêu trên cùng với bảng ước tính số lượng từ vựng trẻ em của các nghiên cứu trước đây, chúng tôi đã tiến hành phân tích tương quan và rút ra được dự tính vốn từ vựng cốt lõi của trẻ em người Việt như trong Hình 3 dưới đây. Cũng cần lưu ý rằng, các con số ở đây là ước toán. Thêm vào đó, số từ vựng cốt lõi có thể có những khoảng chung cho đa số trẻ em, nhưng cũng có những khoảng từ vựng mang tính khác biệt cá nhân giữa các trẻ. Thiết nghĩ, đây là một điều bình thường trong ngôn ngữ, bởi lẽ, điều này cũng hiện diện trong đặc trưng ngôn ngữ của người trưởng thành, như nhiều công trình nghiên cứu đã từng công bố.



Hình 3. Số lượng từ theo độ tuổi (thích ứng theo Reynolds & Fletcher-Janzen, 2013)

Danh sách các từ và các thông tin liên quan được trích đặc từ kho ngữ liệu đầy đủ tại đường link sau:

<https://www.dropbox.com/s/zlfv4qiiu51b9n/danh%20s%C3%A1ch%20t%E1%BB%AB%20t%E1%BA%A1o%20%C4%91%C6%B0%E1%BB%9Dng%20link%20%C4%91%C4%83ng%20b%C3%A1o.xlsx?dl=0>

5. Thảo luận

Có nhiều kho ngữ liệu khác nhau phụ thuộc vào nguồn dữ liệu, thể loại dữ liệu cũng như phương pháp thống kê, phân tích, ngày nay với sự trợ giúp của công nghệ thông tin và các phần mềm giúp cho công việc xây dựng kho ngữ liệu có kích thước lớn hơn và đỡ tốn nhiều công sức. Trong nghiên cứu của tác giả Nguyễn Đức Dân đã sử dụng nhiều nguồn dữ liệu trong giai đoạn 1956-1972 bao gồm tác phẩm văn học, báo chí, thơ, kịch, bài viết của Hồ Chí Minh dành cho người lớn và chỉ có 48.500 từ trên tổng số 524.500 từ được tập hợp từ những truyện cho trẻ em, phương pháp thủ công đã đưa ra được công trình rất đồ sộ thời kì đó. Trong nghiên cứu của tác giả Phạm Giang có tập hợp dữ liệu gồm 279 tác phẩm trong nước và 78 tác phẩm in ở nước ngoài dành cho trẻ mầm non đến lớp 5 gồm thể loại truyện tranh, dân gian, ngụ ngôn, trong giai đoạn 1976-2006 có sử dụng phần mềm để thống kê phân tích.

Trong nghiên cứu này chúng tôi lựa chọn tất cả dữ liệu từ những nhà xuất bản trong nước với những thể loại truyện tranh, truyện dân gian, ngụ ngôn, chỉ có 2 cuốn sách tiếng Việt tập 1 và tập 2 xuất bản năm 2018 là sách giáo khoa thống nhất dùng chung cho toàn quốc, trong việc lựa chọn có tiêu chí là các sách truyện này phải phù hợp với trẻ 6 tuổi do vậy có thể đó là những truyện cho trẻ mầm non dưới 6 tuổi hoặc tài liệu cho trẻ lớn hơn nhưng phải bao gồm trẻ 6 tuổi, điều này giúp xây dựng các từ thông dụng nhất với tần suất xuất hiện cao cho trẻ 6 tuổi, từ cơ sở đó để xây dựng danh sách các từ thủ ngữ Việt để đo thính lực lời cho trẻ em tuổi học đường từ 6-15 tuổi. Phân tích từ loại chỉ ra rằng danh từ gồm danh từ chung và danh từ chiếm 40,6%, động từ chiếm 34,3%, tính từ chiếm 18,3%, kết quả phân tích này cũng phù hợp với nghiên cứu của Phạm Giang với tỉ lệ 43,2% danh từ, 36,7% động từ, 23,1% tính từ.

Chúng tôi trích rút 1.000 từ có tần suất xuất hiện cao nhất trong kho ngữ liệu để phân tích, so sánh đồng thời làm cơ sở để xây dựng các danh sách từ thủ đo thính lực lời về sau được dễ dàng, thuận lợi. Có 799 từ đơn tiết trong 1.000 từ chiếm 79,9% kết quả này cũng tương ứng với tỉ lệ 70,1% từ đơn tiết trong 1.000 từ có tần suất xuất hiện cao nhất trong kho ngữ liệu của tác giả Nguyễn Đức Dân. Khi phân tích sự trùng lặp trong số 799 từ 1 âm tiết của chúng tôi thì có tới 513 từ trùng lặp với nghiên cứu của tác giả Nguyễn Đức Dân và 666 từ trùng lặp với nghiên cứu của tác giả Phạm Giang, qua đây chúng ta thấy các từ thông dụng, tần suất cao có tỷ lệ trùng lặp giữa các nghiên cứu cao mặc dù dữ liệu xây dựng kho ngữ liệu khác nhau về thể loại cũng như thời điểm nghiên cứu. Tuy vậy, vẫn cần xây dựng kho ngữ liệu riêng đặc biệt là cho trẻ em phục vụ cho việc đánh giá sát thực nhất cho nhóm lứa tuổi này.

Về tính thoả đáng ngữ liệu trong nghiên cứu này, theo chúng tôi, có liên quan đến vai trò của văn học đối với việc thụ đắc ngôn ngữ trẻ em. Vai trò của văn học đối với trẻ em đã nhận được sự chú ý của các nhà nghiên cứu về phát triển năng lực ngôn ngữ của trẻ. Văn học thể hiện nhu cầu của con người trong việc truyền đạt ý nghĩa, và thường được gắn liền với bối cảnh xã hội hoặc văn hóa. Do đó, bản thân văn học bao gồm ý nghĩa văn hóa xã hội quan trọng. Người viết văn học thiếu nhi chủ yếu là người lớn. Mục đích của nó nói chung là để giải trí và hướng dẫn trẻ em, và văn học thiếu nhi thường mang một số hình thức chủ đề hoặc đạo đức.

Văn học thiếu nhi về cơ bản rất quan trọng đối với sự phát triển ngôn ngữ. Trẻ em không chỉ được hưởng lợi từ việc tự đọc mà các kĩ năng ngôn ngữ của chúng cũng phát triển đáng kể khi nghe người lớn đọc truyện cho. Văn học ảnh hưởng đến sự phát triển ngôn ngữ của trẻ em chủ yếu bằng cách cung cấp một mô hình để đọc, viết và nói. Thông qua việc cung cấp một lối thoát cho việc học tập tham gia, trẻ em có cơ hội mở rộng các kĩ năng ngôn ngữ của mình một cách giàu trí tưởng tượng và sáng tạo. Nếu không có sự giới thiệu và mở rộng vốn từ vựng thông qua văn học, trẻ sẽ phải cố gắng nhiều để phát triển vốn từ vựng của riêng mình. Văn học tạo cơ hội cho trẻ em làm điều đó một cách giàu trí tưởng tượng và thú vị. Các tác giả sử dụng các đặc trưng của ngôn ngữ để thu hút trẻ em và

phát triển các kỹ năng của riêng chúng. Thông qua việc sử dụng các đặc trưng ngôn ngữ này, và ngoài ra thông qua các tình huống đọc độc lập, được chia sẻ, mô hình hóa và có hướng dẫn, văn học cải thiện không chỉ vốn từ vựng của trẻ em mà còn cả cách phát âm của chúng. Điều này được thể hiện rõ trong các văn bản như trong nhiều văn bản dành cho trẻ em với tính văn điệu cao. Việc sử dụng vần điệu và nhịp điệu cũng được thể hiện rõ trong văn bản *Bài thơ yêu mẹ* của tác giả Nguyễn Bao. Hãy xem xét văn bản sau:

*Mẹ đi làm
Từ sáng sớm
Dậy thổi cơm
Mua thịt cá*

*Em kể má
Được mẹ yêu
Ơi mẹ ơi
Con yêu mẹ lắm.*

Có thể nói rằng văn bản này là một ví dụ điển hình về tầm quan trọng của minh họa trong văn học thiếu nhi và tầm quan trọng của nó trong phát triển ngôn ngữ. Thông qua các hình ảnh minh họa, trẻ có thể bắt đầu tạo ra sự kết nối giữa từ ngữ và hình ảnh, đặc biệt rõ ràng trong văn bản này là việc sử dụng hình minh họa để mô tả các tính từ. Điều này cực kỳ có lợi cho sự phát triển ngôn ngữ. Đó là, sự phát triển ngôn ngữ của độc giả trẻ mang lại lợi ích nhiều hơn thông qua ứng dụng và kinh nghiệm hơn là chỉ đơn giản là lắng nghe.

Thông qua việc sử dụng văn học thiếu nhi, sự phát triển ngôn ngữ của độc giả nhỏ tuổi được tăng cường thông qua việc đọc, viết và nghe. Cho dù trẻ em tự đọc, hay trải nghiệm văn học thông qua các trải nghiệm đọc được chia sẻ hoặc mô hình hóa, văn học của trẻ em đều cực kì quan trọng trong sự phát triển ngôn ngữ của chúng. Giáo viên và cha mẹ trẻ có thể đọc hoặc kể chuyện cho trẻ nghe. Trong khi làm việc đó, người lớn cần truyền đạt một cách tự do và diễn cảm để trẻ nắm được hết các nghĩa nội ngôn và ngoại ngôn của văn bản. Giáo viên hoặc cha mẹ trẻ cần nắm chắc nội dung cơ bản của chuyện, có thể giải thích thêm cho trẻ, và có thể sử dụng từ mới khi truyền đạt.

Trong một nghiên cứu gần đây, nhóm tác giả Umek và các cộng sự (2007) đã ghi nhận tác động của việc đọc văn học của trẻ em đối với sự phát triển ngôn ngữ ở trẻ mẫu giáo. Các tác giả này cho thấy tầm quan trọng của vai trò của văn học trẻ em đối với sự phát triển tâm lí, xã hội và ngôn ngữ của trẻ và trong việc phát triển các kĩ năng học tập cơ bản của trẻ, chẳng hạn như đọc và viết, đã được xác nhận bởi nhiều nghiên cứu. Một vấn đề trung tâm trong tâm lí học phát triển là những hoạt động liên quan đến sách thiếu nhi có ảnh hưởng đến sự phát triển của trẻ và theo những cách nào. Theo các tác giả này, mối quan tâm này đối với sách thiếu nhi và phát triển ngôn ngữ trẻ em đặt nghiên cứu của họ vào hai ngành khoa học là tâm lí học và ngôn ngữ học. Nghiên cứu khám phá tác động của việc đọc có hệ thống và thường xuyên các cuốn sách dành cho trẻ em được chọn trong các cơ sở giáo dục mầm non đối với sự phát triển năng lực ngôn ngữ ở trẻ em từ bốn đến sáu tuổi, ở cả hai giới.

6. Kết luận

Trong bài viết này, chúng tôi đã trình bày việc xây dựng kho ngữ liệu có kích thước 176.153 từ. Xuất phát từ kho ngữ liệu trên, chúng tôi tính toán và đo đạc được 9.573 lượt từ được tập hợp từ dữ liệu 232 truyện, sách dành cho trẻ em góp phần bổ sung vào kho ngữ liệu tiếng Việt đang có; Đưa ra được Danh sách 1.000 từ ngữ có tần suất xuất hiện nhiều nhất trong kho ngữ liệu ngôn ngữ trẻ em. Dựa trên khối từ vựng trên, chúng tôi đã tiến hành trích đặc từ vựng cốt lõi cho trẻ em nói tiếng Việt từ 1 đến 6 tuổi dựa trên phương pháp thống kê trong ngôn ngữ học ngữ liệu theo số lượng từ vựng biểu đạt như trình bày trong Hình 3. Đồng thời, kết quả nghiên cứu này có thể được dùng làm cơ sở để nghiên cứu ứng dụng trong việc xây dựng danh sách từ phổ thông, từ cốt lõi trong lĩnh vực ngôn ngữ cũng như xây dựng danh sách từ thử ứng dụng trong y học để đánh giá khả năng nghe hiểu lời nói của cơ quan thính giác.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Nguyễn Đức Dân, Đặng Thái Minh (2000), *Thống kê ngôn ngữ học - Một số ứng dụng*. Nxb Giáo dục.
2. Nguyễn Đức Dân (1980), *Dictionnaire de fréquence du vietnamien*. Université de Paris VII.
3. Hoàng Phê (2020), *Từ điển tiếng Việt*. Nxb Đà Nẵng.

4. Đặng Thái Minh (1999), *Từ điển điện tử tần số tiếng Việt (Với các tiện ích phục vụ ngôn ngữ học so sánh)*. Tóm tắt luận án Tiến sĩ Ngữ văn, Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Thành phố Hồ Chí Minh.
5. Đoàn Thiện Thuật (2007), *Ngữ âm tiếng Việt*. Nxb Đại học Quốc gia Hà Nội.

Tiếng Anh

6. Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993), Vocabulary Development: A Morphological Analysis. *Monographs of the Society for Research in Child Development*, 58(10), i. doi:10.2307/1166112
7. Ho, Laina. (2000), Children's literature in adult education. *Children's Literature in Education* 31 (4): 259–271.
8. Lecours, A. R. (1975), *Myelogenetic correlates of the development of speech and language*. In E. H. Lenneberg, & E. Lenneberg (Eds.), *Foundations of language development: A multidisciplinary approach* (Vol. 1, pp. 121–135). New York: Academic Press.
9. Lestrade S (2017), Unzipping Zipf's law. *PLoS ONE* 12(8): e0181987. <https://doi.org/10.1371/journal.pone.0181987>
10. Lê, H. P., Nguyen, T. M. H., Roussanaly, A., & Ho, V. (2008), A hybrid approach to word segmentation of Vietnamese texts. In C. Martin-Vide, F. Otto, & H. Fernau (Eds.), *Language and automata theory and applications* (Vol. 5196, pp. 240–249)., Lecture Notes in Computer Science Springer: Berlin, Heidelberg.
11. Lê, H. P., Roussanaly, A., Nguyen, T. M. H., & Rossignol, M. (2010), An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts. In *Traitement Automatique des Langues Naturelles—TALN 2010* (p. 12), Montréal Canada. ATALA (Association pour le Traitement Automatique des Langues).
12. Ljubica Marjanovi Umek, Urška Fekonja, Simona Kranjc & Petra Lešnik Musek (2003), The impact of reading children's literature on language development in the preschool child, *European Early Childhood Education Research Journal*, 11:1, 125-135, DOI: 10.1080/13502930385209111
13. Massi, Maria and Adriana Benvenuto. (2001), Using fairy tales to develop reading and writing skills. *CATESOL Journal* 13: 157–164.
14. Owens, R. E., Jr. (1984), *Language development: An introduction*. Columbus, OH: Charles E. Merrill Publishing.
15. Pham, G., Kohnert, K., & Carney, E. (2008), Corpora of Vietnamese Texts: Lexical Effects of Intended Audience and Publication Place. *Behavior Research Methods*, 40, 154-163.
16. Pham, H., Tucker, B.V. & Baayen, R.H. (2019), Constructing two Vietnamese corpora and building a lexical database. *Lang Resources & Evaluation* 53, 465–498. <https://doi.org/10.1007/s10579-019-09451-x>
17. Reynolds, C. R., & Fletcher-Janzen, E. (2013), *Handbook of Clinical Child Neuropsychology*. Springer US.
18. Segbers, Jutta and Schroeder, Sascha. (2017), How many words do children know? A corpus-based estimation of children's total vocabulary size, *Language Testing*, Vol. 34(3) 297–320.

Measuring the child vocabulary using corpus linguistic methods

Abstract: Developing a corpus for children is a necessary job to serve as a basis for linguistics as well as in the medical field to create the words lists for speech audiometry to assess children's ability to listen and understand speech. Using the method of corpus linguistics, a corpus of 176153 words has been built from the data of 232 stories and books for children. Gauging the correlations between the frequencies of the current corpus with some available Vietnamese corpora, we estimate the vocabulary size of children (from 1 to 6 year olds). The 1000 words with the highest frequency in the corpus are also analyzed to serve as a basis for future research.

Key words: corpus; child language; word frequency; lexical dispersion; Vietnamese.