Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00013

ĐÁNH GIÁ TÍNH DỄ ĐỌC CỦA VĂN BẢN TIẾNG VIỆT DỰA TRÊN WORDNET

Phạm Duy Tâm, Trần Minh Hùng, Lương An Vinh, Đinh Điền

Trung tâm Ngôn ngữ học Tính toán - Trường ĐH Khoa học Tự nhiên Tp. Hồ Chí Minh

1212346@student.hcmus.edu.vn, 1212157@student.hcmus.edu.vn, anvinhluong@gmail.com, ddien@fit.hcmus.edu.vn

TÓM TẮT— *Tính dễ đọc của một văn bản là tổng hợp các yếu tố của văn bản tác động tới khả năng đọc và hiểu hoàn toàn nội dung của văn bản. Việc đánh giá tính dễ đọc có vai trò rất lớn trong quá trình soạn thảo văn bản nhằm xác định đúng đối tượng độc giả muốn hướng đến. Những nghiên cứu về tính dễ đọc của văn bản đã được thực hiện từ lâu trên thế giới nhưng chủ yếu là cho tiếng Anh và một số ngôn ngữ phổ biến khác,... Đối với tiếng Việt, đã có 2 công trình nghiên cứu về vấn đề này nhưng chỉ thực hiện trên các đặc trưng bề mặt của ngôn ngữ như độ dài từ, độ dài câu,... Trong bài báo này, chúng tôi tiến hành thực nghiệm lại một phương pháp đánh giá tính dễ đọc của văn bản dựa trên bộ từ điển ngữ nghĩa WordNet cho tiếng Anh và tiến hành một số thay đổi để thực nghiệm trên bộ WordNet tiếng Việt. Những kết quả đạt được cho thấy đây là một phương pháp tiềm năng và có thể sử dụng làm cơ sở cho các nghiên cứu sau này về đánh giá tính dễ đọc văn bản cho tiếng Việt.*

Từ khóa— *Tính dễ đọc của văn bản – text readablity, từ điển ngữ nghĩa WordNet.*

I. GIỚI THIỆU

Tính dễ đọc của văn bản (text readability) – theo định nghĩa của Edgar Dale và Jeanne Chall (1949) [7] là "tổng hợp các yếu tố của một văn bản ảnh hưởng đến sự thành công của một nhóm người đọc văn bản đó. Sự thành công ở đây là mức độ họ hiểu văn bản đó, đọc nó với một tốc độ tối ưu và cảm thấy thích thú khi đọc văn bản đó". Tính dễ đọc thường nhầm lẫn với tính dễ nhìn (legibility) của văn bản là "mức độ dễ dàng đọc của một văn bản dựa trên các yếu tố như kiểu chữ, kích cỡ chữ, khoảng cách dòng,…". Tính dễ đọc của văn bản có tác động rất lớn tới khả năng đọc và hiểu hoàn toàn văn bản. Căn cứ vào tính dễ đọc của văn bản, người đọc có thể xác định được văn bản mình muốn đọc có phù hợp với khả năng của mình hay không. Người tạo ra văn bản cũng có thể căn cứ vào tính dễ đọc của văn bản đang soạn thảo để định hướng đối tượng người đọc hay có những điều chỉnh cho phù hợp hơn với đối tượng người đọc đang hướng tới.

Việc xây dựng được một mô hình để phân tích tính dễ đọc của văn bản có ý nghĩa rất lớn trong khoa học và thực tiễn: giúp các nhà khoa học có thể viết các báo cáo nghiên cứu dễ đọc hơn cho đối tượng người đọc đang hướng tới; hỗ trợ các nhà giáo dục soạn thảo các sách giáo khoa, giáo trình phù hợp với từng lứa tuổi và trình độ của học sinh, sinh viên; hỗ trợ các nhà xuất bản trong việc định hình đối tượng độc giả; giúp các cơ quan soạn thảo văn bản quy phạm pháp luật có thể điều chỉnh được nội dung cho phù hợp với đa số công dân; hay giúp các nhà sản xuất trong việc soạn thảo các tài liệu hướng dẫn sử dụng các sản phẩm của họ,… Ngoài ra, việc xác định được tính dễ đọc của văn bản có thể hỗ trợ rất hiệu quả trong việc lựa chọn giáo trình phù hợp khi giảng dạy ngôn ngữ cho người nước ngoài.

Trên thế giới đã có rất nhiều các công trình nghiên cứu về việc xác định và phân loại tính dễ đọc của văn bản và hầu hết đều là cho tiếng Anh. Từ giữa thế kỉ XIX, đã có một các số khảo sát về khả năng đọc viết của người trưởng thành ở Mỹ tiêu biểu là khảo sát của Louis Harris [11], nghiên cứu của Khảo sát Tiến bộ Giáo dục Quốc gia (National Assessment of Educational Progress – NAEP) [11],... Các kết quả của các cuộc khảo sát đã thể hiện được sự khác biệt lớn của kỹ năng đọc viết ở người lớn và mức độ ảnh hưởng của khả năng đọc viết đến cuộc sống.

Cuối thế kỉ XIX, đã có rất nhiều công thức đánh giá tính dễ đọc của văn bản được đề xuất, một số công thức phổ biến có thể kể đến như công thức tính dễ đọc Flesch [6], Dale-Chall [7],… Các công thức trên chủ yếu sử dụng các đặc trưng đơn giản ở mức bề mặt văn bản như độ dài từ, độ dài câu, từ vựng,… Cũng trong giai đoạn này, các nhà xuất bản, nhà giáo dục và người giảng dạy bắt đầu quan tâm đến việc sử dụng các công thức đánh giá tính dễ đọc của văn bản để hỗ trợ cho việc lựa chọn văn bản, tài liệu cho phù hợp với người đọc, người học…

Bắt đầu từ những năm 50 của thế kỉ XX, việc đánh giá tính dễ đọc của văn bản đã có những bước phát triển mới. Các nghiên cứu ở giai đoạn này đã bắt đầu đề cập đến sự đóng góp của các yếu tố tâm lý học như sở thích, động lực và kiến thức của cá nhân ảnh hưởng đến tính dễ đọc nhằm tăng độ chính xác của việc đánh giá tính dễ đọc của văn bản.

Năm 2008, nhóm tác giả Shu-yen Lin [10] cùng các cộng sự có một công trình nghiên cứu về phương pháp đánh giá tính dễ đọc của văn bản tiếng Anh dựa trên bộ từ điển ngữ nghĩa WordNet[1]. Họ đã sử dụng mối quan hệ ngữ nghĩa trên WordNet như hạ danh (hyponyms), thượng danh (hypernyms) để xác định các từ cơ bản (basic word) để đánh giá tính dễ đọc của văn bản. Trong bài báo này, chúng tôi tiến hành thực nghiệm lại phương pháp này trên một bộ ngữ liệu văn bản tiếng Anh.

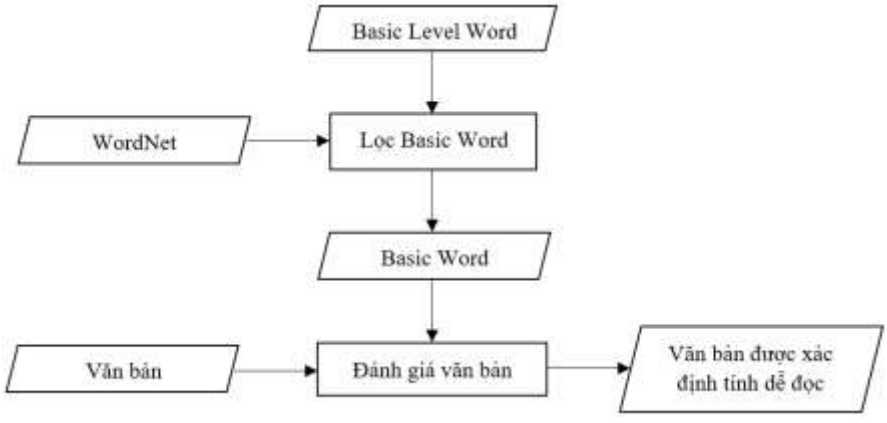
[1] WordNet là một cơ sở dữ liệu tri thức ngữ nghĩa từ vựng được xây dựng theo hệ thống cấp bậc. Hệ thống cấp bậc trong WordNet được xác định bằng các mối quan hệ ngữ nghĩa giữa các từ vựng.

Đối với tiếng Việt, cho tới hiện giờ chỉ có hai công trình nghiên cứu về tính dễ đọc của văn bản tiếng Việt của nhóm tác giả Liem Thanh Nguyen và Alan B. Henkin (năm 1982 và 1985) [1, 2] thực hiện cho cộng đồng người Việt ở hải ngoại. Hai nghiên cứu này tập trung vào mối liên hệ giữa các đặc điểm thống kê cấp độ từ và câu với tính dễ đọc của văn bản trên một tập ngữ liệu nhỏ (24 văn bản) chứ chưa đi vào phân tích các đặc điểm sâu hơn như vai trò của từ, ngữ, cấu trúc ngữ pháp, ngữ nghĩa của câu,… Trong bài báo này, chúng tôi cũng tiến hành thực nghiệm phương pháp của nhóm Shu-yen Lin [10] trên một bộ ngữ liệu 10.000 văn bản tiếng Việt dựa trên bộ từ điển ngữ nghĩa WordNet tiếng Việt cùng với một số thay đổi cho phù hợp. Các kết quả thực nghiệm cho thấy đây là một phương pháp tiềm năng và có thể sử dụng làm cơ sở cho việc nghiên cứu sau này đối với vấn đề đánh giá tính dễ đọc của văn bản cho tiếng Việt.

Phần tiếp theo của bài báo sẽ mô tả chi tiết về phương pháp thực nghiệm. Kết quả thực nghiệm và kết luận sẽ lần lượt được trình bày ở Phần 3 và Phần 4.

II. PHƯƠNG PHÁP

Hình 1 mô tả kiến trúc của hệ thống phương pháp của nhóm tác giả Shu-yen Lin [10]. Đầu tiên, tập danh sách các từ mức độ cơ bản (Basic Level Word – BLW) sẽ được lọc lại thông qua WordNet để xác định lại danh sách các từ cơ bản (Basic Word – BW). Khái niệm BLW, theo định nghĩa của Rosch [9], là những từ thường dễ tiếp nhận hơn các từ hạ danh (hyponyms) và thượng danh (hypernyms). Thượng danh là một quan hệ ngữ nghĩa trong WordNet, là từ có lớp ngữ nghĩa bao hàm từ khác (ví dụ: ‘màu sắc’ sẽ là thượng danh của ‘màu đỏ’). Tương tự, hạ danh là một từ có ngữ nghĩa cụ thể trong tập con của từ có lớp ngữ nghĩa rộng hơn (ví dụ ‘màu đỏ’ sẽ là hạ danh của ‘màu sắc’). Phương pháp của nhóm Shu-yen Lin [10] chỉ thực nghiệm trên từ loại danh từ. Tiếp theo, danh sách BW đã được lọc sẽ được dùng để đánh giá tính dễ đọc của văn bản đưa vào.



Hình 1. Mô hình kiến trúc hệ thống của phương pháp đánh giá tính dễ đọc của văn bản tiếng Anh dựa trên bộ từ điển ngữ nghĩa WordNet của nhóm tác giả Shu-yen Lin.

A. Lọc BW

1. Thực nghiệm 1: Thống kê độ dài và độ phức tạp của hạ danh và thượng danh

Mục tiêu của thực nghiệm là khảo sát độ dài và độ phức tạp của các BLW và các từ thuộc hạ danh và thượng danh trực tiếp của BLW trên WordNet. BLW được giả định có các đặc trưng sau: độ dài từ tương đối ngắn (bao gồm ít ký tự hơn độ dài trung bình của các từ thuộc hạ danh và thượng danh); hạ danh trực tiếp có nhiều tập đồng nghĩa (synsets)[2] hơn thượng danh trực tiếp; hình thái từ đơn giản. Tập các BLW của Rosch [9] được thống kê về độ dài, độ phức tạp và số tập đồng nghĩa ở mỗi BLW và hạ danh, thượng danh của nó. Các kết quả thống kê là cơ sở cho việc xác nhận các giả định về tính chất BLW đặt ra ở đầu thực nghiệm. Các kết quả thực nghiệm sẽ được trình bày ở phần 3, mục A.

2. Thực nghiệm 2: Thống kê tỉ lệ BLW trong cấu tạo từ ghép của hạ danh

Mục tiêu của thực nghiệm là khảo sát sự đóng góp của BLW, hạ danh và thượng danh trực tiếp của BLW trong cấu tạo từ ghép. Nhóm Shu-yen Lin giả định rằng BLW tham gia cấu tạo nên các từ ghép nhiều hơn hạ danh và thượng danh trực tiếp của nó. Với mỗi BLW trong thực nghiệm 1 cùng với các từ thuộc hạ danh và thượng danh trực tiếp, nhóm Shu-yen Lin thống kê tất cả từ trong hạ danh của từ đang xét và các từ ghép mà từ đang xét tham gia cấu tạo, nhằm thống kê tỉ lệ số từ ghép của hạ danh mà từ đang xét tham gia cấu tạo trên tất cả từ hạ danh. Từ ghép là từ được cấu tạo từ hai từ đơn lẻ trở lên (ví dụ: ‘thiếu nữ’ là từ ghép được cấu tạo bởi hai từ đơn). Đối với các từ có nhiều hơn một nhánh nghĩa, phương pháp chỉ tập trung nhánh nghĩa theo định nghĩa của Rosch [9]. Các kết quả thống kê là cơ sở cho việc xác nhận các giả định về tính chất BLW đặt ra ở đầu thực nghiệm. Các kết quả thực nghiệm sẽ được trình bày ở Phần 3, mục A.

[2] Tập đồng nghĩa (synsets) là tập hợp các từ và cụm từ đồng nghĩa với nhau (ví dụ: táo sẽ có hạ danh là hai tập đồng nghĩa cây táo và trái táo).

3. Hai điều kiện lọc

Dựa trên các kết quả sơ bộ của hai thực nghiệm, nhóm Shu-yen Lin giả định BW sẽ có hai tính chất: (1) nó xuất hiện nhiều trong các từ ghép hạ danh; (2) chiều dài từ ngắn hơn chiều dài trung bình của các hạ danh trực tiếp. Các tính chất trên có thể đơn giản thành điều kiện lọc để xác định BW:

(1) Tỉ lệ từ ghép của tất cả hạ danh ≥ 25%;

(2) Độ dài trung bình của hạ danh trực tiếp trừ độ dài từ đang xét ≥ 4.

Dựa trên hai tính chất và điều kiện lọc, thông tin cần thiết để mỗi danh từ xác định có phải là BW bao gồm (1) độ dài từ đó (số ký tự của từ); (2) tỉ lệ từ ghép của từ đó (số từ ghép của hạ danh mà từ đó tham gia cấu tạo); (3) Độ dài trung bình của hạ danh trực tiếp. Kết quả thống kê danh sách BW đã lọc sẽ được trình bày ở Phần 3, mục A.

B. Đánh giá mối liên hệ giữa BW và tính dễ đọc của văn bản

Mục tiêu của thực nghiệm là đánh giá mối liên hệ giữa BW và tính dễ đọc của văn bản. Nhóm Shu-yen Lin giả định một văn bản dễ đọc sẽ chứa nhiều BW hơn văn bản khó đọc hơn; nghĩa là, tỉ lệ BW trong văn bản dễ đọc sẽ cao hơn văn bản khó đọc hơn. Nhằm đảm bảo tính khách quan, nhóm Shu-yen Lin tiến hành đánh giá mối liên hệ giữa BW và tính dễ đọc trên một tập các văn bản thông qua thống kê tỉ lệ BW trên tổng số danh từ ở mỗi văn bản, các văn bản đã được đánh giá tính dễ đọc bằng phương pháp khác. Các kết quả thực nghiệm sẽ được trình bày ở Phần 3, mục A.

C. Đánh giá mối liên hệ giữa BW và tính dễ đọc trên văn bản tiếng Việt

Chúng tôi cũng áp dụng phương pháp của nhóm Shu-yen Lin để thực nghiệm trên tiếng Việt. Các mục tiêu, giả định và phương pháp ở mỗi thực nghiệm vẫn được thực hiện tương tự nhưng sẽ có một số thay đổi cho phù hợp với tiếng Việt. Đầu tiên, chúng tôi sẽ sử dụng WordNet tiếng Việt của Trung tâm Ngôn ngữ học tính toán[3] - Trường Đại học Khoa học Tự nhiên Thành phố Hồ Chí Minh. Tiếp theo, việc đánh giá mối liên hệ giữa BW và tính dễ đọc của văn bản của phương pháp sẽ được tiến hành trên bộ ngữ liệu tiếng Việt tự xây dựng với các mức độ tính dễ đọc khác nhau do chúng tôi giả định, nội dung này sẽ được trình bày chi tiết ở Phần 3, mục B.

III. THỰC NGHIỆM

Ở phần này, chúng tôi sẽ lần lượt trình bày các kết quả thực nghiệm trên phương pháp của nhóm Shu-yen Lin trên cả tiếng Anh và tiếng Việt.

A. Các kết quả thực nghiệm trên tiếng Anh

1. Thực nghiệm 1: Thống kê độ dài và độ phức tạp của hạ danh và thượng danh

Để đạt được mục tiêu của thực nghiệm này, chúng tôi đã tiến hành khảo sát trên 4 bộ ngữ liệu gồm 20 từ theo định nghĩa bởi Rosch [9]; 3.000 từ tiếng Anh phổ biến theo thống kê tần số sử dụng[4]; 3.000 danh từ được thống kê trên ngữ liệu Penn Tree Bank [8] và tất cả danh từ thuộc WordNet tiếng Anh. Kết quả thống kê trên 20 từ theo định nghĩa bởi Rosch được trình bày ở Bảng 1. Các kết quả thống kê của các tập ngữ liệu còn lại sẽ lần lượt được trình bày ở Phụ lục 1, 2 và 3 ở cuối bài báo. Bộ từ điển ngữ nghĩa WordNet tiếng Anh online của Đại học Princeton[5] được sử dụng để tiến hành thực nghiệm.

Bảng 1. Kết quả thống kê độ dài (trung bình), số tập đồng nghĩa và độ phức tạp của hình thái* của 20 từ theo định nghĩa Rosh so sánh với hạ danh và thượng danh trực tiếp của nó

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
screwdriver	11	A	8	1	B	20.33	3	B
guitar	6	A	18	1	B	10.33	6	A, B
hammer	6	A	7	1	A	0	0	N/A
piano	5	A	18.67	3	B	10.67	3	A B
apple	5	A	7.5	2	A, B	10.67	3	B
peach	5	A	9	1	B	0	0	N/A
grape	5	A	11	1	B	11.67	3	A, B
pants	5	A	10	1	A	0	0	N/A
socks	5	A	7	1	A	7.4	5	A, B
shirt	5	A	7	1	A	7.667	9	A, B
table	5	A	5	1	A	13	6	A, B
chair	5	A	4	1	A	11.33	15	A, B
truck	5	A	12	1	B	8.455	11	A, B

[3] Computional Linguistics Center – CLC. Website: http://www.clc.hcmus.edu.vn

[4] http://www.wordfrequency.info/free.asp

[5] http://wordnet.princeton.edu

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
drum	4	A	20	1	B	7.5	8	A, B
lamp	4	A	20	1	B	10	17	A, B
saw	3	A	6	1	A	0	0	N/A
car	3	A	12	1	B	7.968	31	A, B
bus	3	A	15	1	B	8.667	3	A, B
dog	3	A	10	2	A, B	7.5	18	A, B
cat	3	A	6	1	A	9	2	A, B

*A đại diện cho từ đơn, B đại diện cho từ ghép

Các kết quả thống kê đã xác nhận các giả định ban đầu của thực nghiệm. Đầu tiên, độ dài trung bình từ của cả hạ danh và thượng danh đều lớn hơn BLW. Thứ hai, hạ danh có nhiều tập đồng nghĩa hơn thượng danh. Thứ ba, BLW luôn có hình thái đơn giản. Ngược lại, từ hạ danh và thượng danh trực tiếp có hình thái phức tạp hơn. Một số từ hạ danh là từ ghép. Thượng danh phức tạp hơn, mỗi BLW (trừ 'peach') có ít nhất một từ ghép thượng danh của nó.

2. Thực nghiệm 2: Thống kê tỉ lệ BLW trong cấu tạo từ ghép của hạ danh

Bảng 2 trình bày kết quả thống kê trên 20 từ theo định nghĩa bởi Rosch [9]. Tương tự, các kết quả thống kê của các tập ngữ liệu còn lại sẽ lần lượt trình bày ở Phụ lục 4, 5 và 6 ở cuối bài báo.

Bảng 2. Kết quả thống kê tỉ lệ từ ghép và sự phân phối từ ghép trong các cấp độ hạ danh của 20 từ theo định nghĩa Rosch

Thượng danh BLW Hạ danh	Số từ ghép/ Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
hand tool	0/241	0	0	0	0	0	0	0
screwdriver	4/4	100	0	4	0	0	0	0
flat tip screwdriver	0/0	N/A	0	0	0	0	0	0
stringed instrument	1/85	2	0	1	0	0	0	0
guitar	5/12	42	0	5	0	0	0	0
acoustic guitar	0/0	N/A	0	0	0	0	0	0
striker	0/0	N/A	0	0	0	0	0	0
hammer	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
keyboard instrument	0/35	0	0	0	0	0	0	0
piano	8/16	50	0	4	4	0	0	0
grand piano	3/8	38	0	3	0	0	0	0
edible fruit	0/303	0	0	0	0	0	0	0
apple	5/29	18	0	5	0	0	0	0
cooking apple	0/4	0	0	0	0	0	0	0
fruit tree	2/450	1	0	2	0	0	0	0
peach	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
edible fruit	0/303	0	0	0	0	0	0	0
grape	6/17	36	0	3	2	1	0	0
muscadine	0/0	N/A	0	0	0	0	0	0
underpants	0/17	0	0	0	0	0	0	0
pants	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
hosiery	0/30	0	0	0	0	0	0	0

Thượng danh BLW Hạ danh	Số từ ghép/ Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
socks	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
anklet	0/0	N/A	0	0	0	0	0	0
garment	4/445	1	0	3	1	0	0	0
shirt	8/17	48	0	8	0	0	0	0
camise	0/0	N/A	0	0	0	0	0	0
array	1/49	3	0	1	0	0	0	0
table	7/10	70	0	6	1	0	0	0
actuarial table	0/1	0	0	0	0	0	0	0
seat	0/1	0	0	0	0	0	0	0
chair	31/48	65	0	17	13	1	0	0
armchair	0/10	0	0	0	0	0	0	0
motor vehicle	0/153	0	0	0	0	0	0	0
truck	15/48	32	0	10	5	0	0	0
dump truck	0/0	N/A	0	0	0	0	0	0
percussion instrument	0/68	0	0	0	0	0	0	0
drum	5/14	36	0	5	0	0	0	0
bass drum	0/0	N/A	0	0	0	0	0	0
source of illumination	0/107	0	0	0	0	0	0	0
lamp	27/68	40	0	19	7	1	0	0
aladdin's lamp	0/0	N/A	0	0	0	0	0	0
saying	0/59	0	0	0	0	0	0	0
saw	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
motor vehicle	0/153	0	0	0	0	0	0	0
car	21/76	28	0	19	2	0	0	0
ambulance	0/1	0	0	0	0	0	0	0
public transport	0/38	0	0	0	0	0	0	0
bus	3/5	60	0	3	0	0	0	0
minibus	0/0	N/A	0	0	0	0	0	0
canine	0/1	0	0	0	0	0	0	0
dog	51/279	19	0	11	22	16	2	0
basenji	0/0	N/A	0	0	0	0	0	0
feline	0/123	0	0	0	0	0	0	0
cat	35/87	41	0	4	30	1	0	0
domestic cat	0/32	0	0	0	0	0	0	0

Từ Bảng 2 cho thấy, hầu hết mỗi BLW đều có tỉ lệ tham gia cấu tạo từ ghép được là thống kê cao nhất. So sánh với hạ danh và thượng danh của BLW, nó được sử dụng nhiều trong cấu tạo từ ghép. Tuy nhiên, có một số từ (ví dụ: 'crab apple') có tỉ lệ thống kê cao nhưng không được cho là BLW vì các từ ghép mà nó tham gia cấu tạo từ ghép đã bao gồm BLW (ví dụ: 'Southern crab apple'), trường hợp trên đại diện cho tính chết kế thừa của các từ ghép có cấu tạo từ BLW.

3. Kết quả thống kê tỉ lệ trung bình của BW trên ngữ liệu tiếng Anh

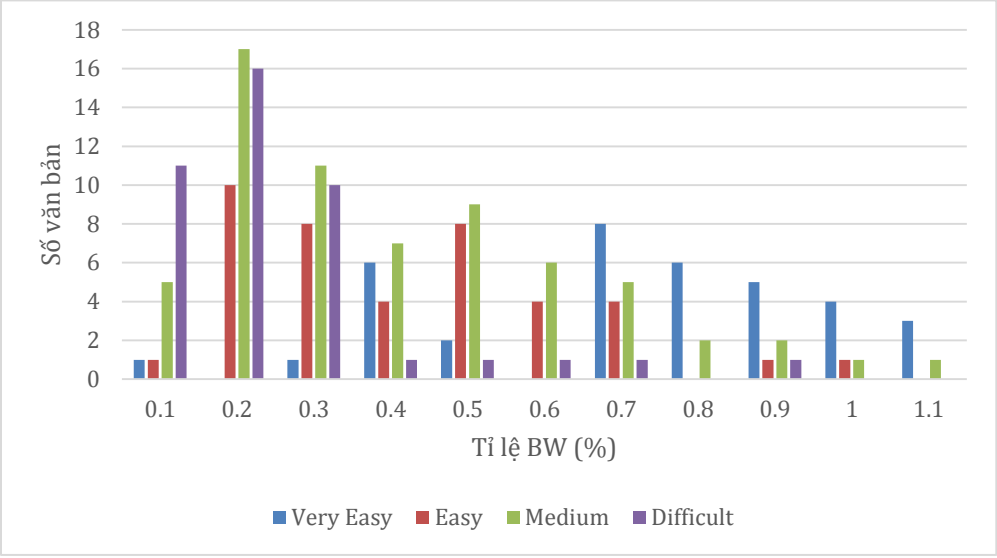
Dựa trên kết quả thống kê của hai thực nghiệm trên, nhóm Shu-yen Lin đã rút ra điều kiện lọc cho việc xác định BW đã được trình bày ở Phần 2. Kết quả lọc BW gồm 13 BW trên 20 từ theo định nghĩa Rosch, 294 BW trên

3.000 từ tiếng Anh phổ biến theo thống kê tần số sử dụng, 389 BW trên 3.000 danh từ được thống kê trên ngữ liệu Penn Tree Bank và 2.505 BW trên tất cả danh từ thuộc WordNet tiếng Anh. Trong bài báo này, chúng tôi đã tiến hành thực nghiệm đánh giá tính dễ đọc trên bộ ngữ liệu sách giáo khoa tiếng Anh của nhóm Islam [5]. Bộ ngữ liệu này bao gồm 519 văn bản, 95.470 câu và 1.184.124 từ theo định dạng TEI P5. Các kết quả thống kê dựa trên cả 4 bộ ngữ liệu BW và các kết quả đánh giá tính dễ đọc bằng 2 công thức Flesh Grade Level [6] và Dale-Chall [7] được trình bày ở Bảng 3.

Bảng 3. Tỉ lệ trung bình BW và kết quả đánh giá tính dễ đọc bằng 2 công thức Flesh Grade Level, Dale-Chall trên mỗi cấp độ khác nhau của ngữ liệu tiếng Anh

Cấp độ	Tỉ lệ BW				Flesh Grade Level	Dale-Chall
	20 từ theo định nghĩa của Rosch	3.000 từ phổ biến	3.000 danh từ phổ biến	Tất cả danh từ WordNet		
1	0.359	7.488	7.969	14.766	4.569	6.742
2	0.156	6.655	7.375	14.324	5.608	6.907
3	0.165	6.572	7.494	14.602	6.571	6.975
4	0.103	5.878	7.006	14.264	7.760	7.053

Các văn bản này được chia thành 4 cấp độ. Cấp độ 1 (level 1) được giả định là dễ nhất, cấp độ 4 (level 4) là khó nhất. Kết quả thống kê ở Bảng 3 và Hình 2 thể hiện tỉ lệ BW giảm theo độ khó của văn bản. Tỉ lệ BW ở các văn bản cấp độ 1 nhiều hơn cấp độ 4. Ta cũng có thể thấy tỉ lệ BW trung bình ở cấp độ 2 thường gần xấp xỉ với cấp độ 3, nguyên nhân có thể do độ khó của các văn bản thuộc 2 cấp độ này không chênh lệch nhiều lắm. Chúng tôi sẽ khảo sát kĩ hơn về nguyên nhân trong các nghiên cứu sau này.



Hình 2. Thống kê số lượng văn bản theo từng mức tỉ lệ BW trên ngữ liệu tiếng Anh với tập 20 từ của Rosch

B. Các kết quả thực nghiệm trên tiếng Việt

Tương tự với phương pháp của nhóm Shu-yen Lin trên tiếng Anh, bài báo này cũng tiến hành các thực nghiệm trên tiếng Việt và thay đổi ngữ liệu cho phù hợp. Đối với thực nghiệm 1 và 2, chúng tôi đã tiến hành khảo sát trên 3 bộ ngữ liệu gồm 3000 từ phổ biến theo thống kê tần số sử dụng [3]; 3000 danh từ phổ biến theo thống kê tần số sử dụng [3] và tất cả danh từ thuộc WordNet tiếng Việt. Đối với thực nghiệm 3, bộ ngữ liệu tiếng Việt do chúng tôi tự xây dựng được sử dụng để tiến hành thực nghiệm. Chúng tôi đã xây dựng bộ ngữ liệu tiếng Việt với 3 cấp độ tính dễ đọc khác nhau. Cấp độ đầu tiên – cấp độ dễ (easy level) được thu thập từ sách giáo khoa (từ lớp 2 cho đến lớp 5); truyện ngắn dành cho thiếu nhi; văn mẫu; các tin tức trên các website thiếu nhi. Cấp độ thứ hai – cấp độ trung bình (normal level) được thu thập từ các website tin tức hằng ngày như Dân trí, Tuổi trẻ, Thanh niên, VnExpress, Vietnamnet,… Cấp độ cuối cùng – cấp độ khó (difficult level) được thu thập từ Các bài viết trên các tạp chí lý luận về Đảng, Nhà nước; các bài viết trên các tạp chí lý luận về ngôn ngữ, văn hóa, xã hội; các bài luận, giáo trình về Triết học; các văn bản Quy phạm pháp luật,... Các ngữ liệu đều được chúng tôi thu thập từ tài nguyên Internet và được tiền xử lý (tách từ, tách câu, chuẩn hóa văn bản,…) trước khi đưa vào thực nghiệm.

Các kết quả thực nghiệm 1, 2 và 3 theo phương pháp của nhóm Shu-yen Lin trên tiếng Việt được trình bày lần lượt ở các Bảng 4, 5, 6 và Phụ lục 7, 8, 9, 10 ở cuối bài báo. Trong Bảng 6, chúng tôi cũng trình bày kết quả đánh giá tính dễ đọc bằng công thức của nhóm Liem Thanh Nguyen và Alan B. Henkin [1].

Bảng 4. Kết quả thống kê độ dài (trung bình), số tập đồng nghĩa và độ phức tạp của hình thái của 3.000 danh từ tiếng Việt phổ biến theo thống kê tần số sử dụng so sánh với hạ danh và thượng danh trực tiếp của nó

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
ổ	1	A	5	1	B	2	1	A
ô	1	A	10	1	B	0	0	N/A
âm	2	A	14	1	B	20	1	B
tủ	2	A	7	1	B	12	4	B
nón	3	A	8	1	B	9.68	22	A B
trí	3	A	8	1	B	8.8	5	B
suối	4	A	8	1	B	10	6	A B
tiệc	4	A	9	1	B	15.25	16	A B
đường	5	A	9	1	B	10	11	A B
thuốc	5	A	11	1	B	10.67	6	B
xã hội	6	B	12	1	B	10.75	4	B
bí mật	6	B	7	1	B	0	0	N/A
máy ảnh	7	B	12	1	B	9	8	B
binh sĩ	7	B	17	1	B	12.64	14	B
toán học	8	B	7	1	B	14.5	2	B
thể chất	8	B	12	1	B	10.33	3	B
quê hương	9	B	15	1	B	15.64	11	A B
chuyên môn	10	B	14	1	B	12.1	20	B
cổ động viên	12	B	13	1	B	15.5	24	A B
kinh tế thị trường	18	B	13	1	B	13	1	B
…	…	…	…	…	…	…	…	…

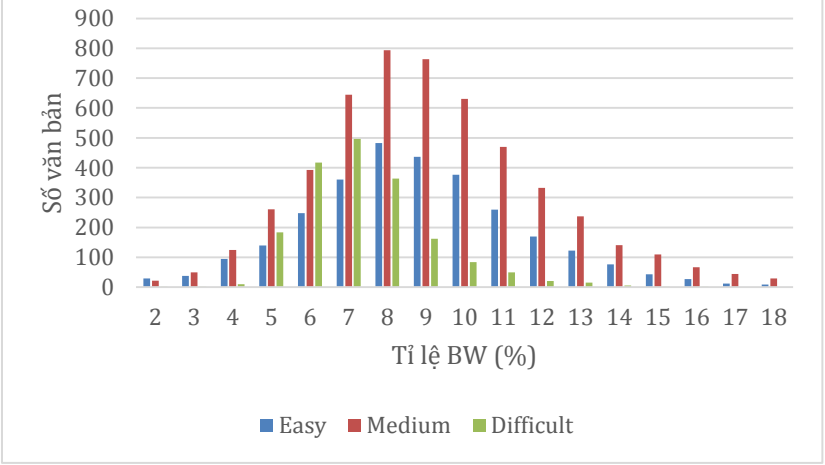
Bảng 5. Kết quả thống kê tỉ lệ từ ghép và sự phân phối từ ghép trong các cấp độ hạ danh của của 3.000 danh từ tiếng Việt phổ biến theo thống kê tần số sử dụng

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
kết tụ	0/833	0	0	0	0	0	0	0
ổ	0/2	0	0	0	0	0	0	0
bộ	0/0	N/A	0	0	0	0	0	0
vật che nắng	0/16	0	0	0	0	0	0	0
ô	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
hiện tượng cơ học	0/44	0	0	0	0	0	0	0
âm	1/1	100	0	1	0	0	0	0
âm thanh có tần số siêu âm	0/0	N/A	0	0	0	0	0	0
kho chứa	0/25	0	0	0	0	0	0	0
tủ	5/8	63	0	4	1	0	0	0
gian để chổi	0/0	N/A	0	0	0	0	0	0
đồ đội đầu	0/196	0	0	0	0	0	0	0
nón	8/73	11	0	5	3	0	0	0
mũ	0/0	N/A	0	0	0	0	0	0
nhận thức	8/216	4	0	4	4	0	0	0
trí	0/7	0	0	0	0	0	0	0
bất tỉnh	0/0	N/A	0	0	0	0	0	0
vùng nước	2/206	1	0	2	0	0	0	0
suối	9/40	23	0	7	2	0	0	0
lạch	0/0	N/A	0	0	0	0	0	0
buổi tụ họp	1/129	1	0	0	1	0	0	0
tiệc	39/69	57	0	23	15	1	0	0
buổi chiêu đãi	0/0	N/A	0	0	0	0	0	0
đại phân tử	0/918	0	0	0	0	0	0	0

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
đường	42/167	26	0	10	9	4	11	8
deoxiriboza	0/0	N/A	0	0	0	0	0	0
phương thuốc	2/85	3	0	2	0	0	0	0
thuốc	4/14	29	0	3	1	0	0	0
bột seidlitz	0/0	N/A	0	0	0	0	0	0
tập đoàn xã hội	0/3838	0	0	0	0	0	0	0
xã hội	5/36	14	0	5	0	0	0	0
văn minh	15/23	66	0	11	4	0	0	0
giấu kín	0/13	0	0	0	0	0	0	0
bí mật	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
trang thiết bị	3/1036	1	0	1	2	0	0	0
máy ảnh	12/79	16	0	1	10	1	0	0
bảng dập báo hiệu	0/0	N/A	0	0	0	0	0	0
công nhân có tay nghề	0/1062	0	0	0	0	0	0	0
binh sĩ	1/362	1	0	0	0	1	0	0
binh nhì	0/0	N/A	0	0	0	0	0	0
khoa học	29/851	4	0	7	8	5	5	4
toán học	4/94	5	0	2	1	1	0	0
toán học thuần tuý	0/64	0	0	0	0	0	0	0
đặc tính cơ thể	0/216	0	0	0	0	0	0	0
thể chất	0/19	0	0	0	0	0	0	0
dáng béo lùn	0/0	N/A	0	0	0	0	0	0
người chủ xướng	0/18	0	0	0	0	0	0	0
cổ động viên	1/110	1	0	0	1	0	0	0
chỗ dựa chính	0/0	N/A	0	0	0	0	0	0
hệ thống kinh tế	0/31	0	0	0	0	0	0	0
kinh tế thị trường	0/4	0	0	0	0	0	0	0
chủ nghĩa tư bản	0/1	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…

Bảng 6. Tỉ lệ trung bình BW và kết quả đánh giá tính dễ đọc bằng công thức của nhóm Liem Thanh Nguyen trên mỗi cấp độ khác nhau của ngữ liệu tiếng Việt

Cấp độ	Tỉ lệ BW			Công thức của nhóm Liem Thanh Nguyen
	3.000 từ phổ biến	3.000 danh từ phổ biến	Tất cả danh từ WordNet	
1	13.343	9.077	15.074	6.377
2	12.585	9.176	14.232	8.832
3	10.196	7.319	11.857	12.738



Hình 3. Thống kê số lượng văn bản theo từng mức tỉ lệ BW trên ngữ liệu tiếng Việt với 3.000 danh từ tiếng Việt phổ biến

Từ Bảng 6 và Hình 3 có thể thấy nhận định về tỉ lệ BW trong văn bản tiếng Việt cũng tương đồng với trong tiếng Anh: độ khó tăng thì tỉ lệ BW giảm. Đối với kết quả trên tập 3.000 danh từ phổ biến, tỉ lệ BW trung bình trên cấp độ 1 và 2 gần xấp xỉ nhau, nguyên nhân sẽ được chúng tôi khảo sát nguyên nhân trong các nghiên cứu kế tiếp.

IV. KẾT LUẬN

Trong bài báo này, chúng tôi đã tiến hành thực nghiệm lại phương pháp đánh giá độ khó của văn bản của nhóm tác giả Shu-yen Lin trên một bộ ngữ liệu tiếng Anh lớn hơn và tiến hành một số thay đổi để thực nghiệm trên ngữ liệu tiếng Việt. Các kết quả thực nghiệm trên bộ ngữ liệu tiếng Anh và tiếng Việt đã xác nhận lại nhận định là văn bản càng khó thì tỉ lệ từ cơ bản càng ít. Tuy vẫn còn một số kết quả không thực sự rõ ràng nhưng tổng quan là văn bản dễ sẽ có ít từ cơ bản hơn văn bản khó. Trong các nghiên cứu tiếp theo, chúng tôi sẽ tiến hành khảo sát trên tập ngữ liệu tiếng Việt lớn hơn và mở rộng sang các từ loại khác (như động từ, tính từ,…) chứ không chỉ là danh từ như trong nghiên cứu này.

TÀI LIỆU THAM KHẢO

[1] B. H. Liem T. Nguyen, "A Second Generation Readability Formula for Vietnamese", Journal of Reading, vol. 29, pp. 219-225, 1985.

[2] B. H. Liem Thanh Nguyen, "A Readability Formula for Vietnamese", Journal of Reading, vol. 26, pp. 243-251, 1982.

[3] Đinh Điền, Đỗ Đức Hào, "Chữ Quốc ngữ hiện nay qua các con số thống kê", Hội thảo Chữ Quốc ngữ - Phú Yên.

[4] E. Dale, J. S. Chall, "A formula for predicting readability", Educational research bulletin, vol. 27, no.1, pp. 11-20, 28, 1948.

[5] Islam, M. Zahurul, "Multilingual Text Classification using Information - Theoretic Features", PhD Thesis-Goethe University Frankfurt, 2014.

[6] J. N. Farr, J. J. Jenkins, D. G. Paterson. "Simplification of the Flesch Reading Ease Formula", Journal of applied psychology, vol. 35, no. 5, pp. 333-357, 1951.

[7] J. S. C. Edgar Dale, "The Concept of Readability", Elementary English, vol. 26, pp. 19-26, 1949.

[8] M. Marcus, B. Santorini, M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank", Computational Linguistics - Special issue on using large corpora: II, vol. 19, no. 2, pp. 313-330, 1993.

[9] Rosch, Eleanor, Mervis, Carolyn, Gray, Wayne, Johnson, David, & Boyes-Braem, Penny, "Basic objects in natural categories", Cognitive Psychology, vol. 8, pp. 382-439, 1976.

[10] S. Y. Lin, C.C. Su, Y. D. Lai, L.C. Yang, S.K Hsieh, "Measuring Text Readability by Lexical Relations Retrieved from Wordnet", Proceedings of the 20th Conference on Computational Linguistics and Speech Processing, 2008.

[11] T. G. Sticht, A. B. Armstrong, "Adult literacy in the United States: A compendium of quantitative data and interpretive comment", 1994.

ASSESSING VIETNAMESE TEXT READABILITY USING WORDNET

Pham Duy Tam, Tran Minh Hung, Luong An Vinh, Dinh Dien

ABSTRACT— *Text readability is a combination of factors in a text that affects its reading comprehension. Assessing text readability plays an important role in text creating process, which helps to deliver the right content to right target readers. Although text readability has been studied for a long time, proposed researches mainly focus on English and other popular languages. In Vietnamese, there have been two studies using shallow features like word length and sentence length. In this paper, we conduct an experiment on text readability measurement based on English WordNet and Vietnamese WordNet with adjustment. The results show that this is a potential method which could be used as fundamental for future researches on Vietnamese text readability.*

V. PHỤ LỤC

Phụ lục 1. Kết quả thống kê độ dài (trung bình), số tập đồng nghĩa và độ phức tạp của hình thái của 3.000 từ tiếng Anh phổ biến theo thống kê tần số sử dụng so sánh với hạ danh và thượng danh trực tiếp của nó

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
concentration	13	A	8	1	A	11.25	4	A, B
fundamental	11	A	6	1	A	0	0	N/A
maintenance	11	A	6	1	A	11.33	6	A, B
extension	9	A	5	1	A	0	0	N/A
listening	9	A	7	1	A	11.5	2	A
childhood	9	A	10	1	B	7.5	2	A
distances	9	A	7	1	A	7.94	18	A, B
landscape	9	A	7	1	A	0	0	N/A
miracle	7	A	9	1	A	0	0	N/A
mission	7	A	27	1	B	0	0	N/A
airport	7	A	8	1	A	8	1	A
battles	7	A	14	1	B	9.17	6	A, B
help	4	A	8	1	A	7.33	12	A
does	4	A	19	1	B	0	0	N/A
jean	4	A	7.5	2	A	6	1	A
data	4	A	10	1	A	9.67	3	A, B
bee	3	A	19	1	B	10.3	10	A, B
fan	3	A	6	1	A	9	3	A, B
joy	3	A	7	1	A	9.75	4	A
in	2	A	10	1	B	0	0	N/A
…	…	…	…	…	…	…	…	…

Phụ lục 2. Kết quả thống kê độ dài (trung bình), số tập đồng nghĩa và độ phức tạp của hình thái của 3.000 danh từ tiếng Anh được thống kê trên ngữ liệu Penn Tree Bank so sánh với hạ danh và thượng danh trực tiếp của nó

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
telecommunications	18	A	6	1	A	9.6	5	A
recommendations	15	A	6	1	A	8	1	A
developers	10	A	7	1	A	0	0	N/A
permission	10	A	8	1	A	7	8	A, B
difficulty	10	A	6	1	A	7	2	A, B
comparison	10	A	11	1	A	9	5	A
assistant	9	A	6	1	A	9.52	27	A, B
hospitals	9	A	15	1	B	12	6	A, B
behavior	8	A	8	1	A	10.86	7	A, B
cities	6	A	12	1	A	14.67	3	B
flight	6	A	9	1	A	0	0	N/A
values	6	A	6	1	A	0	0	N/A
rumors	6	A	6	1	A	0	0	N/A
lesson	6	A	8	1	A	11.6	5	B
track	5	A	4	1	A	7.67	6	A, B
genes	5	A	8	1	A	11	21	A, B
cells	5	A	11	1	A	0	0	N/A
bias	4	A	10	1	A	11.86	7	A, B
lake	4	A	11	1	B	5.75	8	A, B
bike	4	A	12	1	B	8.5	2	A, B
…	…	…	…	…	…	…	…	…

Phụ lục 3. Kết quả thống kê độ dài (trung bình), số tập đồng nghĩa và độ phức tạp của hình thái của tất cả danh từ WordNet tiếng Anh so sánh với hạ danh và thượng danh trực tiếp của nó

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
remark	6	A	9	1	A	8.33	12	A B
reserve	7	A	9	1	A	10	1	A
promise	7	A	9.5	2	A B	7	5	A B
obsession	9	A	16	1	B	12	1	A
poker face	10	B	11	1	A	0	0	N/A
open house	10	B	5	1	A	0	0	N/A
motor home	10	B	19	1	B	3	1	A
home movie	10	B	5	1	A	0	0	N/A
door prize	10	B	5	1	A	0	0	N/A
cross hair	10	B	4	1	A	0	0	N/A
club member	11	B	6	1	A	0	0	N/A
car company	11	B	7	1	A	0	0	N/A
livingroom set	14	B	5	1	A	0	0	N/A
kitchen police	14	B	14	1	B	0	0	N/A
imperial beard	14	B	5	1	A	0	0	N/A
greenwich time	14	B	4	1	A	24	1	B
flowering cherry	16	B	6	1	A	13.5	4	A B
first appearance	16	B	9	1	A	13	4	A B
chemical engineering	20	B	11	1	A	0	0	N/A
basketball backboard	20	B	13	1	B	0	0	N/A
…	…	…	…	…	…	…	…	…

Phụ lục 4. Kết quả thống kê tỉ lệ từ ghép và sự phân phối từ ghép trong các cấp độ hạ danh của 3.000 từ tiếng Anh phổ biến theo thống kê tần số sử dụng

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
property	7/123	6	0	7	0	0	0	0
concentration	3/14	22	0	3	0	0	0	0
hydrogen ion concentration	0/6	0	0	0	0	0	0	0
factor	3/10	30	0	2	1	0	0	0
fundamental	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
repair	1/47	3	0	0	1	0	0	0
maintenance	1/10	10	0	1	0	0	0	0
camera care	0/0	N/A	0	0	0	0	0	0
delay	0/3	0	0	0	0	0	0	0
extension	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
sensing	0/0	N/A	0	0	0	0	0	0
listening	1/6	17	0	1	0	0	0	0
auscultation	0/3	0	0	0	0	0	0	0
time of life	0/64	0	0	0	0	0	0	0

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
childhood	0/4	0	0	0	0	0	0	0
boyhood	0/0	N/A	0	0	0	0	0	0
spacing	0/0	N/A	0	0	0	0	0	0
distances	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
elevation	0/0	N/A	0	0	0	0	0	0
scenery	0/12	0	0	0	0	0	0	0
landscape	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
happening	0/1718	0	0	0	0	0	0	0
miracle	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
nongovernmental organization	0/65	0	0	0	0	0	0	0
mission	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
airfield	1/11	10	0	1	0	0	0	0
airport	0/1	0	0	0	0	0	0	0
heliport	0/0	N/A	0	0	0	0	0	0
military action	0/101	0	0	0	0	0	0	0
battles	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
armageddon	0/0	N/A	0	0	0	0	0	0
activity	9/5223	1	0	3	2	2	2	0
help	4/89	5	0	2	2	0	0	0
accommodation	0/4	0	0	0	0	0	0	0
executive department	0/61	0	0	0	0	0	0	0
does	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
trouser	2/49	5	0	2	0	0	0	0
jean	0/2	0	0	0	0	0	0	0
levi's	0/0	N/A	0	0	0	0	0	0
collection	6/576	2	0	4	2	0	0	0
data	3/3	100	0	3	0	0	0	0
accounting data	0/0	N/A	0	0	0	0	0	0
hymenopterous insect	0/92	0	0	0	0	0	0	0
bee	18/26	70	0	11	7	0	0	0
andrena	0/0	N/A	0	0	0	0	0	0
device	29/4466	1	0	18	9	2	0	0
fan	3/5	60	0	2	1	0	0	0
electric fan	0/0	N/A	0	0	0	0	0	0
emotion	3/268	2	0	3	0	0	0	0
joy	0/22	0	0	0	0	0	0	0
elation	0/9	0	0	0	0	0	0	0
linear unit	2/147	2	0	2	0	0	0	0
in	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
…	…	…	…	…	…	…	…	…

Phụ lục 5. Kết quả thống kê tỉ lệ từ ghép và sự phân phối từ ghép trong các cấp độ hạ danh của 3.000 danh từ tiếng Anh được thống kê trên ngữ liệu Penn Tree Bank

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
medium	4/102	4	0	2	2	0	0	0
telecommunications	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
broadcasting	0/0	N/A	0	0	0	0	0	0
advice	1/11	10	0	1	0	0	0	0
recommendations	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
referral	0/0	N/A	0	0	0	0	0	0
creator	0/6	0	0	0	0	0	0	0
developers	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
approval	0/26	0	0	0	0	0	0	0
permission	0/20	0	0	0	0	0	0	0
all clear	0/0	N/A	0	0	0	0	0	0
effort	0/152	0	0	0	0	0	0	0
difficulty	0/2	0	0	0	0	0	0	0
the devil	0/0	N/A	0	0	0	0	0	0
examination	5/67	8	0	5	0	0	0	0
comparison	0/5	0	0	0	0	0	0	0
analogy	0/0	N/A	0	0	0	0	0	0
worker	24/1467	2	0	6	4	9	4	1
assistant	5/147	4	0	4	1	0	0	0
accomplice	0/4	0	0	0	0	0	0	0
medical building	0/36	0	0	0	0	0	0	0
hospitals	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
creche	0/0	N/A	0	0	0	0	0	0
activity	9/5223	1	0	3	2	2	2	0
behavior	0/46	0	0	0	0	0	0	0
aggression	0/0	N/A	0	0	0	0	0	0
municipality	0/15	0	0	0	0	0	0	0
cities	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
national capital	0/0	N/A	0	0	0	0	0	0
formation	1/55	2	0	1	0	0	0	0
flight	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
belief	2/697	1	0	2	0	0	0	0
values	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
gossip	0/0	N/A	0	0	0	0	0	0
rumors	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
teaching	2/33	7	0	2	0	0	0	0
lesson	10/10	100	0	5	5	0	0	0

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
dance lesson	0/0	N/A	0	0	0	0	0	0
line	4/15	27	0	4	0	0	0	0
track	1/9	12	0	1	0	0	0	0
collision course	0/0	N/A	0	0	0	0	0	0
sequence	1/105	1	0	1	0	0	0	0
genes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
allele	2/4	50	0	2	0	0	0	0
compartment	0/8	0	0	0	0	0	0	0
cells	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
partiality	0/0	N/A	0	0	0	0	0	0
bias	1/11	10	0	1	0	0	0	0
experimenter bias	0/0	N/A	0	0	0	0	0	0
body of water	0/102	0	0	0	0	0	0	0
lake	3/20	16	0	3	0	0	0	0
bayou	0/0	N/A	0	0	0	0	0	0
motor vehicle	0/153	0	0	0	0	0	0	0
bike	4/6	67	0	4	0	0	0	0
minibike	0/1	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…

Phụ lục 6. Kết quả thống kê tỉ lệ từ ghép và sự phân phối từ ghép trong các cấp độ hạ danh của tất cả danh từ WordNet tiếng Anh

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
statement	8/662	2	0	7	0	1	0	0
remark	0/35	0	0	0	0	0	0	0
ad-lib	0/0	N/A	0	0	0	0	0	0
propriety	0/25	0	0	0	0	0	0	0
reserve	0/1	0	0	0	0	0	0	0
demureness	0/0	N/A	0	0	0	0	0	0
commitment	0/1	0	0	0	0	0	0	0
promise	0/11	0	0	0	0	0	0	0
betrothal	0/1	0	0	0	0	0	0	0
irrational motive	0/25	0	0	0	0	0	0	0
obsession	0/1	0	0	0	0	0	0	0
onomatomania	0/0	N/A	0	0	0	0	0	0
countenance	0/11	0	0	0	0	0	0	0
poker face	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
party	4/49	9	0	4	0	0	0	0
open house	0/0	N/A	0	0	0	0	0	0

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
recreational vehicle	0/7	0	0	0	0	0	0	0
motor home	0/2	0	0	0	0	0	0	0
van	0/0	N/A	0	0	0	0	0	0
movie	2/37	6	0	2	0	0	0	0
home movie	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
prize	2/7	29	0	2	0	0	0	0
door prize	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
line	4/15	27	0	4	0	0	0	0
cross hair	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
member	6/40	16	0	6	0	0	0	0
club member	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
company	41/83	50	0	28	4	9	0	0
car company	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
suite	0/1	0	0	0	0	0	0	0
livingroom set	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
enlisted person	0/90	0	0	0	0	0	0	0
kitchen police	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
beard	2/10	20	0	2	0	0	0	0
imperial beard	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
time	0/0	N/A	0	0	0	0	0	0
greenwich time	0/8	0	0	0	0	0	0	0
coordinated universal time	0/6	0	0	0	0	0	0	0
cherry	0/0	N/A	0	0	0	0	0	0
flowering cherry	3/12	26	0	3	0	0	0	0
fuji	0/0	N/A	0	0	0	0	0	0
beginning	0/84	0	0	0	0	0	0	0
first appearance	0/7	0	0	0	0	0	0	0
induction	0/11	0	0	0	0	0	0	0
engineering	3/18	17	0	3	0	0	0	0
chemical engineering	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
game equipment	0/109	0	0	0	0	0	0	0
basketball backboard	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
…	…	…	…	…	…	…	…	…

Phụ lục 7. Kết quả thống kê độ dài (trung bình), số tập đồng nghĩa và độ phức tạp của hình thái của 3.000 từ tiếng Việt phổ biến theo thống kê tần số sử dụng so sánh với hạ danh và thượng danh trực tiếp của nó

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
u	1	A	7	1	B	15	2	B
ổ	1	A	5	1	B	2	1	A
cỗ	2	A	5	1	B	6	2	B
bổ	2	A	8	1	B	0	0	N/A
cãi	3	A	7	1	B	7	1	B
lật	3	A	8	1	B	0	0	N/A
quấn	4	A	15	1	B	0	0	N/A
chôn	4	A	4	1	A	9	1	B
huyết	5	A	3	1	A	0	0	N/A
xưởng	5	A	10	1	B	10.5	4	B
thuyết	6	A	9	1	B	8.5	2	B
phượng	6	A	11	1	B	0	0	N/A
tổ quốc	7	B	15	1	B	15.64	11	A, B
số phận	7	B	8	1	B	9.75	4	B
nối tiếp	8	B	11	1	B	10	2	B
tín dụng	8	B	10	1	B	17	1	B
phong phú	9	B	11	1	B	8	2	B
trái phiếu	10	B	9.5	2	B	16.88	16	B
nghề nghiệp	11	B	14	1	B	12	6	B
chuyên ngành	12	B	8	1	B	0	0	N/A
…	…	…	…	…	…	…	…	…

Phụ lục 8. Kết quả thống kê độ dài (trung bình), số tập đồng nghĩa và độ phức tạp của hình thái của tất cả danh từ WordNet tiếng Việt so sánh với hạ danh và thượng danh trực tiếp của nó

Từ / Cụm từ	BLW		Thượng danh			Hạ danh		
	Độ dài	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp	Độ dài	Số tập đồng nghĩa	Độ phức tạp
ù	1	A	10	1	B	14	1	B
ả	1	A	8	1	B	10.5	2	A, B
đô	2	A	11	1	B	10.3913	23	A, B
rê	2	A	6	1	B	0	0	N/A
lốc	3	A	6	1	B	9	1	B
gió	3	A	8	1	B	8.7	30	A, B
ngày	4	A	14	1	B	14	1	B
giàu	4	A	14	1	B	8.4	5	B
tường	5	A	7	1	B	10.8	10	B
giếng	5	A	5	1	B	9	5	B
bạn bè	6	B	3	1	A	10	3	B
cha mẹ	6	B	9	1	B	6.833333	6	A, B
ghế cao	7	B	6	1	B	0	0	N/A
quốc lộ	7	B	8	1	B	9.833333	6	A B
phim ảnh	8	B	8	2	B	9.611111	18	B
phím đàn	8	B	4	1	A	0	0	N/A
hội chứng	9	B	10	1	B	17.76667	30	A, B
điều kỳ lạ	10	B	6	1	B	11	2	B
sự đói khát	11	B	5	1	B	0	0	N/A
kính đeo mắt	12	B	14	1	B	9.6	5	B
…	…	…	…	…	…	…	…	…

Phụ lục 9. Kết quả thống kê tỉ lệ từ ghép và sự phân phối từ ghép trong các cấp độ hạ danh của của 3.000 từ tiếng Việt phổ biến theo thống kê tần số sử dụng

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
kim loại	20/483	5	0	20	0	0	0	0
u	2/2	100	0	2	0	0	0	0
đồng vị Urani 235"	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
kết tụ	0/833	0	0	0	0	0	0	0
ổ	0/2	0	0	0	0	0	0	0
bộ	0/0	N/A	0	0	0	0	0	0
kết tụ	0/833	0	0	0	0	0	0	0
cỗ	1/5	20	0	1	0	0	0	0
bộ bài	0/0	N/A	0	0	0	0	0	0
chia tách	0/12	0	0	0	0	0	0	0
bổ	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
sự hồi âm	0/60	0	0	0	0	0	0	0
cãi	1/11	10	0	1	0	0	0	0
láo xược	0/0	N/A	0	0	0	0	0	0
sự đắm tàu	0/5	0	0	0	0	0	0	0
lật	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
thay đổi hình dạng	0/0	N/A	0	0	0	0	0	0
quấn	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
giấu	5/55	10	0	4	1	0	0	0
chôn	0/1	0	0	0	0	0	0	0
sự cải táng	0/0	N/A	0	0	0	0	0	0
máu	0/0	N/A	0	0	0	0	0	0
huyết	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
khu đất rộng	0/312	0	0	0	0	0	0	0
xưởng	0/6	0	0	0	0	0	0	0
khu chứa đồ cũ	0/0	N/A	0	0	0	0	0	0
học thuyết	0/7	0	0	0	0	0	0	0
thuyết	0/7	0	0	0	0	0	0	0
quy tắc vàng	0/0	N/A	0	0	0	0	0	0
cây gỗ hạt kín	0/1044	0	0	0	0	0	0	0
phượng	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
khu vực hành chính	0/152	0	0	0	0	0	0	0
tổ quốc	1/60	2	0	1	0	0	0	0
các nước chiếm đại lục nam mỹ	0/0	N/A	0	0	0	0	0	0
cuộc diện	0/230	0	0	0	0	0	0	0

số phận	0/163	0	0	0	0	0	0	0
phong vân	0/32	0	0	0	0	0	0	0
sắp xếp thứ tự	0/159	0	0	0	0	0	0	0
nối tiếp	0/9	0	0	0	0	0	0	0
luân phiên	0/0	N/A	0	0	0	0	0	0
tiền cho vay	0/34	0	0	0	0	0	0	0
tín dụng	0/4	0	0	0	0	0	0	0
khoản cho vay mua ô tô	0/0	N/A	0	0	0	0	0	0
nguồn tài sản	0/34	0	0	0	0	0	0	0
phong phú	0/20	0	0	0	0	0	0	0
kho báu	0/0	N/A	0	0	0	0	0	0
chứng khoán	18/84	22	0	15	2	1	0	0
trái phiếu	29/40	73	0	25	3	1	0	0
công tác không có tiền	0/0	N/A	0	0	0	0	0	0
nhóm đồng nghiệp	0/28	0	0	0	0	0	0	0
nghề nghiệp	0/25	0	0	0	0	0	0	0
giới doanh nhân	0/0	N/A	0	0	0	0	0	0
ngành học	0/1	0	0	0	0	0	0	0
chuyên ngành	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
…	…	…	…	…	…	…	…	…

Phụ lục 10. Kết quả thống kê tỉ lệ từ ghép và sự phân phối từ ghép trong các cấp độ hạ danh của của tất cả danh từ WordNet tiếng Việt

Thượng danh BLW Hạ danh	**Số từ ghép / Số hạ danh**	**Tỉ lệ từ ghép (%)**	**Số từ ghép ở mỗi cấp độ hạ danh**					
			Cấp độ 1	**Cấp độ 2**	**Cấp độ 3**	**Cấp độ 4**	**Cấp độ 5**	**Cấp độ 6**
chiến thắng	9/34	27	0	9	0	0	0	0
ù	0/1	0	0	0	0	0	0	0
điểm kết thúc đầu	0/0	N/A	0	0	0	0	0	0
quý nương	0/77	0	0	0	0	0	0	0
ả	1/3	34	0	1	0	0	0	0
lolita	0/0	N/A	0	0	0	0	0	0
đơn vị tiền tệ	0/0	N/A	0	0	0	0	0	0
đô	52/62	84	0	51	1	0	0	0
fiji	0/0	N/A	0	0	0	0	0	0
nốt son	0/15	0	0	0	0	0	0	0
rê	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
bão gió	0/22	0	0	0	0	0	0	0
lốc	2/2	100	0	2	0	0	0	0
gió lốc nhẹ	0/0	N/A	0	0	0	0	0	0
thời tiết	0/0	N/A	0	0	0	0	0	0
gió	94/111	85	0	56	38	0	0	0
bão	0/0	N/A	0	0	0	0	0	0
khoảng thời gian	0/0	N/A	0	0	0	0	0	0

Thượng danh BLW Hạ danh	Số từ ghép / Số hạ danh	Tỉ lệ từ ghép (%)	Số từ ghép ở mỗi cấp độ hạ danh					
			Cấp độ 1	Cấp độ 2	Cấp độ 3	Cấp độ 4	Cấp độ 5	Cấp độ 6
ngày	3/3	100	0	3	0	0	0	0
ngày nguyệt lịch	0/0	N/A	0	0	0	0	0	0
điều kiện kinh tế	0/27	0	0	0	0	0	0	0
giàu	4/26	16	0	2	1	1	0	0
inherited wealth	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
bức vách	0/41	0	0	0	0	0	0	0
tường	16/26	62	0	15	1	0	0	0
tường bao lò	0/0	N/A	0	0	0	0	0	0
hố đào	1/90	2	0	1	0	0	0	0
giếng	13/14	93	0	6	7	0	0	0
giếng dầu	2/7	29	0	2	0	0	0	0
bạn	0/0	N/A	0	0	0	0	0	0
bạn bè	0/9	0	0	0	0	0	0	0
bạn đồng đội	0/0	N/A	0	0	0	0	0	0
cha hoặc mẹ	0/64	0	0	0	0	0	0	0
cha mẹ	3/64	5	0	3	0	0	0	0
cha mẹ kế	0/2	0	0	0	0	0	0	0
ghế tựa	5/78	7	0	5	0	0	0	0
ghế cao	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
con đường	0/0	N/A	0	0	0	0	0	0
quốc lộ	0/22	0	0	0	0	0	0	0
con đường dẫn đến	0/0	N/A	0	0	0	0	0	0
biểu dương	0/197	0	0	0	0	0	0	0
phim ảnh	0/42	0	0	0	0	0	0	0
ba chiều	0/0	N/A	0	0	0	0	0	0
phím	0/0	N/A	0	0	0	0	0	0
phím đàn	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
triệu chứng	0/12	0	0	0	0	0	0	0
hội chứng	28/76	37	0	28	0	0	0	0
a162083z	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
sự kiện	0/2	0	0	0	0	0	0	0
điều kỳ lạ	0/9	0	0	0	0	0	0	0
biến hình	0/0	N/A	0	0	0	0	0	0
sự đói	0/0	N/A	0	0	0	0	0	0
sự đói khát	0/0	N/A	0	0	0	0	0	0
N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
dụng cụ quang học	0/55	0	0	0	0	0	0	0
kính đeo mắt	0/7	0	0	0	0	0	0	0
kính bảo hộ	0/0	N/A	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…