**Vietnamese POS Tag (viPOStag)**

Annotation guidelines for VLSP 2022 parsing shared tasks

#### Category

| **STT** | **POS\_old** | **POS\_2022** | **vnPOS** | **enPOS** |
| --- | --- | --- | --- | --- |
|  | N | N | danh từ | Noun |
|  | M | NUM | số từ | Numeral |
|  | L, Q | DET | lượng từ, từ định lượng | Determiner |
|  | V | V | động từ | Verb |
|  | A | ADJ | tính từ | Adjective |
|  | P | PRO | đại từ | Pronoun |
|  | R | ADV | phụ từ | Adverb |
|  | E | PRE | giới từ | Preposition |
|  | Cc | CC | liên từ kết hợp (song song) | Coordinating Conjunction |
|  | C | SC | liên từ phụ thuộc (chính phụ) | Subordinating Conjunction |
|  | T, Aux | PRT | trợ từ, tiểu từ | Particle |
|  | I | I | cảm từ | Interjection, Exclamation |
|  | FW | FW | Từ nguyên dạng tiếng nước ngoài | Foreign words |
|  | CH, SYM | PUNCT | Dấu câu | Punctuation |
|  | CH | SYM | Kí hiệu | Symbol, Character |
|  | X | X | không (hoặc chưa) xác định  (Nhãn này hiện không xuất hiện trong dữ liệu, do mọi dữ liệu đều được gán nhãn từ loại phù hợp với vai trò ngữ pháp của từ/cụm từ trong câu) | Undetermined or Unknown.  (This tag doesn’t appear in the corpus.) |
|  | y | y | Từ viết tắt  (nhãn này được dùng làm nhãn phụ của một từ loại, ví dụ Ny) | Abbreviate/Acronymic  (this tag is used as a suffix of a main POS, e.g. Ny) |
|  | b | b | Từ vay mượn  (nhãn này được dùng làm nhãn phụ của một từ loại, ví dụ Nb) | Borrowed words  (this tag is used as a suffix of a main POS, e.g. Nb) |
|  | G, MWE | Tuỳ theo chức năng  mà phân vào các lớp từ loại cụ thể | - tổ hợp từ, từ tổ  - ngữ cố định  (mỗi đơn vị từ trong các ngữ cố định, tổ hợp từ được gán nhãn từ loại bình thường, cả tổ hợp từ được đánh dấu với nhãn con MWE, ví dụ:  (NP-MWE (N hang) (ADJ cùng) (N ngõ) (ADJ hẻm))  (VP-MWE (V lặp) (V đi) (V lặp) (V lại))  (AP-MWE (ADJ tối) (N mắt) (ADJ tối) (N mũi))  ) | - Group (group of words)  - Multi-word expression  (these kinds of MWE are now labeled as a subcategory of phrases, e.g.  (NP-MWE (N hang) (ADJ cùng) (N ngõ) (ADJ hẻm))  (VP-MWE (V lặp) (V đi) (V lặp) (V lại))  (AP-MWE (ADJ tối) (N mắt) (ADJ tối) (N mũi))  ) |
|  | Z | Z | yếu tố cấu tạo từ  (Nhãn này không xuất hiện trong dữ liệu, các yếu tố cấu tạo từ được ghép với các từ đi kèm để tạo thành 1 đơn vị từ. Ví dụ phó\_phòng được coi là một từ chứ không phải 2 từ.  Ví dụ dữ liệu có sử dụng nhãn này:  Trong từ siêu\_người\_mẫu/N thì/SC siêu/Z có nghĩa là ‘cao vượt lên trên’.  ) | Bound morphemes  (This tag doesn’t appear in the VTB corpus.) |

#### **Subcategory**

| **STT** | **POS\_old** | **POS\_2022** | **vnPOS** | **example** |
| --- | --- | --- | --- | --- |
|  | N | N | danh từ chung (Noun) | - Chỉ từng sự vật: *người, trâu, bò, nhà, máy, xe, đất, đá, v.v.*  - Chỉ từng sự việc riêng lẻ: *cái (tình), cái (tát), cái (vỗ vai), (ngã một) cái, (đùng một) cái...* |
|  | Np | NNP | danh từ riêng  (Proper noun) | - Tên người: *Nguyễn/NNP Văn/NNP Trỗi/NNP, La/NNP Thị/NNP Cẩm/NNP Chướng/NNP, Tôn/NNP Nữ/NNP Ngọc/NNP Ninh/NNP, Quang/NNP Tèo/NNP, Vũ/NNP Nhôm/NNP, Út/NNP Trọc/NNP, Cường/NNP Đô/NNP La/NNP, v.v.*  - Tên địa lí, khu dân cư: *Hà/NNP Nội/NNP, Điện/NNP Biên/NNP Phủ/NNP, TP Hồ/NNP Chí/NNP Minh/NNP, Quận Hai/NNP Bà/NNP Trưng/NNP, Quận 9/NNP, Đường Phạm/NNP Văn/NNP Đồng/NNP, Quốc\_lộ 1A/NNP, Hồ Gươm/NNP, Chùa Một/NNP Cột/NNP, Thôn 2/NNP, Đội 3/NNP, Toà nhà VP3/NNPy, tầng 10/NNP, Ngõ 222E/NNP, Ngách 81/81/NNP, v.v.*  - Tên tổ chức, thương hiệu: *Đảng/N Cộng\_sản/N Việt/NNP Nam/NNP, Viện/N Nghiên\_cứu/V Công\_nghệ/N FPT/NNPy, Ngân\_hàng/N Thương\_mại/N Cổ\_phần/N Á/NNP Châu/NNP, ACB/NNPy, VPBank/NNPy, v.v.*  - Tên sự kiện: *Cách\_mạng/NNP Tháng/NNP Tám/NNP, v.v.* |
|  | Nc | NC | danh từ loại thể  (Classifier noun)  [nhập chung Nc, Ns, Ne thành Nc] | - Chỉ động vật: *con, cái (kiến, cò, vạc), ...*  - Chỉ thực vật: *bông, đoá, ...*  - Chỉ vật vô sinh, theo cấu trúc **Nc + N**; **Num + Nc + N**: *cái* *(cái bàn,* tôi mua *hai cái này), con* *(thuyền, dao), chiếc, cục, cây* *(cột, nến, rơm), lá* *(gan), quả* *(tim, cật), hạt* *(sạn, bụi, mưa, sương), tờ, mẩu, bánh, mảnh, miếng, tảng* *(băng, thịt), cuốn, quyển, cuộn* *(chỉ), pho, ngôi, toà, túp, căn, thửa, suất, ngọn* *(gió), ...*  *-* Chỉ từng cá nhân người, theo cấu trúc **Nc + NNP**: *thằng, ông (thằng Tý, ông Bá)*; **Nc + N chỉ người, Num + Nc + N chỉ người**: *thằng (thằng Tý, thằng bé, ba thằng bé), tên (tên cướp, hai tên cướp), con* *(Liên), cái* *(cái Liên), đứa (trẻ, con\_gái), cô (gái, Lan), ông* *(Tân, giáo), bà* *(Hoa, giáo), chị* *(Lan, công\_nhân), anh* *(Chí, bộ\_đội), đồng\_chí (bí thư), …* CHÚ Ý: ông/N ấy/Pro, bà/N ta/Pro, cô/N ấy/Pro, lão/N ta/Pro  - Chỉ số lượng nhiều nhưng không xác định, theo cấu trúc **Nc + N**; **Num + Nc + N**:  *+ bầy, đàn, toán, tốp, bọn, lũ, bụi, khóm, lùm, giề (bèo), bè (rau), bộ (bàn ghế, phim), dãy (bàn, núi), bụi, buồng (cau), đám (mây), ...*  *+ cột (khói), giàn, luồng (khói, ánh sáng, điện, nước), ngụm, nhúm, xâu, xiên, mớ, lọn, tia (nắng, nước), giọt (nước, sương), dòng (nước, sông), bó, gói, nắm, túm, ...*  *+ chồng, tập, tệp, xấp, thếp, súc (vải, thịt), đụn, đống...* |
|  | Nu | NU | danh từ chỉ đơn vị đo lường, tiền tệ  (Unit Noun) | *- watt, jun, ha, cm, mm, kg, m², m³, cân, yến, tạ, tấn, lít, ...*  *- hào, đồng, xu, quan, đôla, ...*  - *giờ/tiếng, phút, giây, độ\_C*  Gán nhãn Nu cho các từ chỉ đơn vị đo lường và đơn vị tiền tệ. **Không xác định nhãn Nu** cho các từ chỉ dụng cụ đong lường như ca, bò/bơ, đấu, thùng, can... |
|  | Nux | NUX | tổ hợp danh từ chỉ đơn vị mở rộng  (Extended unit noun) | *- lần/phút, lần/p, chu\_kỳ/p, ml/24h, v.v.* |
|  | M | NUM | số từ (Numeral) | *- một, hai, hai\_mươi, hai\_mươi\_hai, hăm\_hai, ba\_lăm, 3, 9, ½ ...*  *- chục, trăm, nghìn, vạn, ức, triệu, tỉ*  *- nhất, nhì, ba, tư, million ...*  *- một\_triệu\_hai\_trăm\_nghìn, 2\_triệu, 12\_triệu\_rưỡi...*  Chú ý:***một*** trong các tổ hợp sau là lượng từ, không phải số từ (không thể thay *hai, ba…* cho *một* được): *không* ***một*** *người ra đón, có* ***một*** *ai đó đã vào đây,* ***một*** *thời gian sau, nằm* ***một*** *xó, từng người* ***một****, cùng* ***một*** *lúc, hơn* ***một*** *nửa....* (xem 8) |
|  | Mx | NUMX | số từ mở rộng  (Extended numeral) | - *10h, ~~9\_giờ\_tối~~, 10h30p, 7-9h,*  *- 300.000đ, 3cm, 10m2, 2m, 50ml, 370C, 100k, 10g25, 15$, £12, 64GB, một\_tấn\_rưỡi, 50%, ba\_phần\_tư, 100\_phần\_100*  *- 9X {người thế hệ những năm 1990}, 2ker {người thế hệ những năm 2000}* |
|  | L, Q | DET | lượng từ, từ định lượng, định từ (Determiner) | - Determiner là các từ trong tiếng Anh như *the/a (một), some (những, vài), every (mỗi), v.v.* đứng trước danh từ để cho biết danh từ đó đang được sử dụng như thế nào. Trong tiếng Việt chú ý gán *một (a/the)* là Det: *không* ***một*** *người ra đón, có* ***một*** *ai đó đã vào đây,* ***một*** *thời gian sau, nằm* ***một*** *xó, từng người* ***một****, cùng* ***một*** *lúc...*  - Gộp các đại từ định lượng *cả, tất\_cả, tất\_thảy, tất\_tật, hết\_thảy, toàn\_bộ, toàn\_thể* vào đây: *các, những, vài, đôi, cặp, dăm, mọi (người), mỗi (lần), mấy, từng (ngày), chút (thời gian), một\_vài, dăm\_ba, mỗi, mỗi\_một, một\_chút, 1\_chút, mọi, nửa (một nửa), ....* |
|  | V | V | động từ (Verb) | *đi, đứng, nằm, ăn, chơi, học, múa, ...* |
|  | V | AUX | trợ động từ, hệ từ LÀ (Auxiliary Verb) | *là (hệ từ), bị, được, cần, phải, nên, muốn, toan, định, …* |
|  | A | ADJ | tính từ (Adjective) | *đẹp, tốt, xấu, to, lớn, nhỏ, dài, ngắn, ...* |
|  | P | PRO | đại từ (Pronoun) | *tao, mày, nó, ấy, bao\_nhiêu, ai, kia, gì, nào, vậy, thế, sao, ...* |
|  | R | ADV | phụ từ (Adverb) | *đã, sẽ, đang, rất, lắm, hết\_sức, cực\_kì, vô\_cùng, có\_lẽ, ...* |
|  | E | PRE | giới từ (Preposition) | *của, để, từ, đến, cho, với (bàn với mọi người, ăn với cá, có trách nhiệm với công việc), ...* |
|  | CC | CC | liên từ phối hợp (Coordinating Conjunction) | *và, hoặc, hay, với (như hình với bóng, anh với tôi cùng đi, đến khổ với nó), …* |
|  | C | SC | liên từ phụ thuộc (Subordinating Conjunction) | *như, nhưng, nếu, thì, vì, nên, …* |
|  | T, Aux | PRT | trợ từ (Particle) | *à, ư, nhỉ, nhé, ...* |
|  | I | I | cảm từ (Interjection) | *ái\_chà, ôi, chao\_ôi, trời\_ơi, haizzz, hihihi, hehe, ...* |
|  | G, X | ADJ | tổ hợp có thính chất như tính từ | Bao gồm cả các thành ngữ có yếu tố không có nghĩa hoặc mơ hồ về nghĩa (như thập\_tử\_nhất\_sinh, đầu\_tắt\_mặt\_tối,...):  *một\_khác, được\_chăng\_hay\_chớ, thập\_tử\_nhất\_sinh…* |
|  | G, X | ADV | tổ hợp có thính chất như phụ từ, bổ nghĩa cho động từ, tính từ | *làm\_gì, có\_vẻ, ít\_nhất, chí\_ít, may\_ra, hoạ\_chăng, hoạ\_may, một\_mình, ngày\_một, ngày\_càng, càng\_ngày\_càng, đến\_nỗi, ai\_ngờ, (học tập) làm\_sao (cho tốt), ra\_bộ, ra\_cái\_vẻ, ra\_cái\_điều, ra\_điều, được\_cái, …* |
|  | G, X | N | tổ hợp có thính chất như danh từ | *địa\_linh\_nhân\_kiệt, vừa\_qua, vừa\_rồi, mới\_đây, mới\_rồi, rồi\_đây, rồi\_mai, rồi\_ra, một\_mặt, mặt\_khác, …* |
|  | G, X | PRT | tổ hợp có thính chất như trợ từ | *nhất\_là, hoá\_ra, té\_ra, thì\_ra, thế\_ra, thật\_ra, thì\_có, hay\_sao, (trông nó) làm\_sao (ấy), (đẹp) làm\_sao, đi\_nữa, đi\_chăng\_nữa* |
|  | G, X | SC | tổ hợp có thính chất như liên từ phụ thuộc | *làm\_sao, may\_mà, chẳng\_hạn, khác\_nào, chẳng\_qua, đã\_đành, đành\_rằng, phải\_chăng, khốn\_nỗi, khổ\_nỗi, khổ\_một\_nỗi, khốn\_một nỗi, …* |
|  | G, X | V | tổ hợp có thính chất như động từ | *tóm\_lại, nói\_tóm\_lại,* *nói\_riêng, nói\_chung, nhìn\_chung, coi\_là, xem\_là, kể\_cả, thay\_vì, (không ai) làm\_sao, …* |
|  | MW | MWE | ngữ cố định/thành ngữ (Multi-word expression) | Gán nhãn thành phần tuỳ theo ý nghĩa của cụm từ cố định, và gán nhãn POS như bình thường cho từng yếu tố cấu thành:  (NP-MWE (N chuột) (V chạy) (N cùng) (N sào)) |
|  | D | D | tổ hợp từ biểu thị ý nghĩa tục tĩu (Dirty word) | - *đù\_má, đù\_mẹ, vãi\_chưởng, đếch, đéo, vãi, ...* |
|  | X | X | tổ hợp không thể xác định từ loại  (Undetermined group) | - Các tổ hợp từ không xác định rõ ý nghĩa, chủ yếu là chữ Hán và phiên âm chữ Hán: *Cử đầu vọng minh nguyệt, Đê đầu tư cố hương* (Ngẩng đầu nhìn trăng sáng, Cúi đầu nhớ cố hương) |
|  | Z | Z | yếu tố cấu tạo từ  (Bound morphemes) | - Các yếu tố đứng độc lập, không tham gia cấu tạo từ: *bất, vô, phi, tổng...* (theo đó *bất, phi, tổng* trong *bất\_hợp\_tác, phi\_nhân\_đạo, tổng\_kiểm\_tra* là tham gia cấu tạo từ, không gán nhãn Z riêng cho chúng) |
|  | y | y | Từ viết tắt (Abbreviation, Acronym) | Viết tắt tên tổ chức, thuật ngữ  *OPEC, WTO, HIV, AK, Ca, 3D, Mig-27, (mạng) 3G, GS., TS., tr., NQ-TW, ...*  *v.v., T/L, T/M, ...* |
|  | b | b | Từ vay mượn (Borrowed words) | - Từ vay mượn tiếng nước ngoài theo đường phiên âm, chuyển tự hoặc để nguyên dạng và được sử dụng như là từ của tiếng Việt thì nhất loạt gán nhãn POS theo dạng Xb (X là từ loại nào đấy, b là viết tắt của từ borrowed).  *Internet/Nb, email/Nb, video/Nb, chat/Vb, sexy/ADJb,…*  - Từ có cấu tạo theo kiểu của tiếng Việt, gồm 1 yếu tố tiếng Việt + 1 yếu tố nước ngoài, thì không dùng nhãn Xb:  *kẽm\_clorua/N, chụp\_CT/V,…* |
|  | FW | FW | Từ nguyên dạng tiếng nước ngoài (Foreign words) | Từ nguyên dạng tiếng nước ngoài được đưa vào văn bản nhưng không được sử dụng phổ biến trong tiếng Việt (tiếng Việt không mượn) thì nhất loạt gán nhãn POS là **FW**. Nhãn này thường được gán cho những từ ngữ, những cụm từ hay câu tiếng nước ngoài (như tiếng Anh chẳng hạn) được trộn mã hay chuyển mã vào tiếng Việt.  *- made, in, Thailand (made in Thailand), anyway, and, or, but, I, love, you, too (I love you too), how, are, you (how are you), …* |
|  | CH, SYM | PUNCT | Dấu câu (Punctuation) | - Các dấu phân tách các chữ (text), các thành phần của câu nhằm làm cho câu văn được rõ ràng, mạch lạc (một số dấu khi cấu tạo công thức thì lại thuộc SYM): . , ; : ? ! … - () “” <> [] {} |
|  | CH | SYM | Kí hiệu (Symbol) | - Các kí hiệu quy ước: @, #, $, %, ^, &, §, ©… (nếu & dùng như “và” thì chú nhãn liên từ CC).  - Các kí hiệu đánh dấu ở đầu mục: *1., A., a), ii., v.v.*  - Các dấu dùng trong công thức toán: +, −, ×, ÷, =, <, >  - Các tổ hợp khác như công thức toán học, hoá học: *x² = a, x+y = 36, √A² = │A│, A1 x √A²*  - Các kí hiệu biểu cảm: :), ♥‿♥, 😝  - Chuỗi chữ - số - kí hiệu biểu thị thông tin gì đó (thường không biết chính xác nguồn gốc chữ viết tắt): 12A1, B2345, 29A.1007, john.doe@universal.org, http://universaldependencies.org/, 1-800-COMPANY |