

Vietnamese POS Tag (viPOSTag)

Annotation guidelines for VLSP 2022 parsing shared tasks

Category

STT	POS_old	POS_2022	vnPOS	enPOS
1.	N	N	danh từ	Noun
2.	M	NUM	số từ	Numeral
3.	L, Q	DET	lượng từ, từ định lượng	Determiner
4.	V	V	động từ	Verb
5.	A	ADJ	tính từ	Adjective
6.	P	PRO	đại từ	Pronoun
7.	R	ADV	phụ từ	Adverb
8.	E	PRE	giới từ	Preposition
9.	Cc	CC	liên từ kết hợp (song song)	Coordinating Conjunction
10.	C	SC	liên từ phụ thuộc (chính phụ)	Subordinating Conjunction
11.	T, Aux	PRT	trợ từ, tiểu từ	Particle
12.	I	I	cảm từ	Interjection, Exclamation
13.	FW	FW	Từ nguyên dạng tiếng nước ngoài	Foreign words
14.	CH, SYM	PUNCT	Dấu câu	Punctuation
15.	CH	SYM	Kí hiệu	Symbol, Character
16.	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.)
17.	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)

18.	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)
19.	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 ngỗ) (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 ngỗ) (ADJ hẻm)) (VP-MWE (V lặp) (V đi) (V lặp) (V lại)) (AP-MWE (ADJ tối) (N mắt) (ADJ tối) (N mũi)))
20.	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
1	N	N	danh từ chung (Noun)	- Chỉ từng sự vật: <i>người, trâu, bò, nhà, máy, xe, đất, đá</i>, v.v. - Chỉ từng sự việc riêng lẻ: <i>cái (tình), cái (tát), cái (vỗ vai), (ngã một) cái, (đùng một) cái...</i>
2	Np	NNP	danh từ riêng (Proper noun)	- Tên người: <i>Nguyễn/NNP Văn/NNP Trỗi/NNP, La/NNP Thị/NNP Cẩm/NNP Chương/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</i>, v.v. - Tên địa lí, khu dân cư: <i>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</i>, v.v. - Tên tổ chức, thương hiệu: <i>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_phân/N Á/NNP Châu/NNP, ACB/NNPy, VPBank/NNPy</i>, v.v. - Tên sự kiện: <i>Cách_mạng/NNP Tháng/NNP Tám/NNP</i>, v.v.
3	Nc	NC	danh từ loại thể (Classifier noun) [nhập chung Nc, Ns, Ne thành Nc]	- Chỉ động vật: <i>con, cái (kiến, cò, vạc)</i>, ... - Chỉ thực vật: <i>bông, đoá</i>, ... - Chỉ vật vô sinh, theo cấu trúc Nc + N; Num + Nc + N: <i>cái (cái bàn, tôi mua hai cái này), con (thuyền, dao), chiếc, cục, cây (cột, nền, rom), 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)</i>, ... - Chỉ từng cá nhân người, theo cấu trúc Nc + NNP: <i>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)</i>, ... 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: + <i>bầy, đàn, toán, tốp, bọn, lũ, bụi, khóm, lùm, giẻ (bèo), bè (rau), bộ (bàn ghế, phim), dãy (bàn, núi), bụi, buồng (cau), đám (mây), ...</i> + <i>cột (khói), giàn, luồng (khói, ánh sáng, điện, nước), ngum, nhúm, xâu, xiên, mớ, lọn, tia (năng, nước), giọt (nước, sương), dòng (nước, sông), bó, gói, nắm, túm, ...</i> + <i>chồng, tập, tềp, xấp, thếp, súc (vải, thịt), đụn, đống...</i>
4	Nu	NU	danh từ chỉ đơn vị đo lường, tiền tệ (Unit Noun)	- <i>watt, jun, ha, cm, mm, kg, m², m³, cân, yến, tạ, tấn, lít, ...</i> - <i>hào, đồng, xu, quan, đôla, ...</i> - <i>giờ/tiếng, phút, giây, độ_C</i> 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 đo lường như ca, xô/bô, chậu, thùng, can...
5	Nux	NUX	tổ hợp danh từ chỉ đơn vị mở rộng (Extended unit noun)	- <i>lần/phút, lần/p, chu_kỳ/p, ml/24h, v.v.</i>
6	M	NUM	số từ (Numeral)	- <i>một, hai, hai_mươi, hai_mươi_hai, hăm_hai, ba_lăm, 3, 9, ½ ...</i> - <i>chục, trăm, nghìn, vạn, ức, triệu, tỉ</i> - <i>nhất, nhì, ba, tư, million ...</i> - <i>một triệu hai trăm nghìn, 2 triệu, 12 triệu rưỡi...</i> Chú ý: <i>một</i> trong các tổ hợp sau là lượng từ, không phải số từ (không thể thay <i>hai, ba...</i> cho <i>một</i> được): <i>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....</i> (xem 8)
7	Mx	NUMX	số từ mở rộng (Extended numeral)	- <i>10h, 9_giờ_tối, 10h30p, 7-9h,</i> - <i>300.000đ, 3cm, 10m², 2m, 50ml, 37°C, 100k, 10g25, 15\$, £12, 64GB, một_tấn_rưỡi, 50%, ba_phần_tư, 100_phần_100</i> - <i>9X {người thế hệ những năm 1990}, 2ker {người thế hệ những năm 2000}</i>
8	L, Q	DET	lượng từ, từ định lượng, định từ (Determiner)	- Determiner là các từ trong tiếng Anh như <i>the/a (một), some (những, vài), every (mỗi)</i>, 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

				<i>một (a/the) là Det: không một người ra đón, có một ai đó đã vào đây, một thời gian sau, năm một xò, từng người một, cùng một lúc...</i> - Gộp các đại từ định lượng <i>cả, tất_cả, tất_thảy, tất_tật, hết_thảy, toàn_bộ, toàn_thể</i> vào đây: <i>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),</i>
9	V	V	động từ (Verb)	<i>đi, đứng, nằm, ăn, chơi, học, múa, ...</i>
10	V	AUX	trợ động từ, hệ từ LÀ (Auxiliary Verb)	<i>là (hệ từ), bị, được, cần, phải, nên, muốn, toan, định, ...</i>
11	A	ADJ	tính từ (Adjective)	<i>đẹp, tốt, xấu, to, lớn, nhỏ, dài, ngắn, ...</i>
12	P	PRO	đại từ (Pronoun)	<i>tao, mày, nó, ấy, bao_nhiêu, ai, kia, gì, nào, vậy, thế, sao, ...</i>
13	R	ADV	phụ từ (Adverb)	<i>đã, sẽ, đang, rất, lắm, hết_sức, cực_kì, vô_cùng, có_lẽ, ...</i>
14	E	PRE	giới từ (Preposition)	<i>của, để, từ, đến, cho, với (bàn với mọi người, ăn với cá, có trách nhiệm với công việc), ...</i>
15	CC	CC	liên từ phối hợp (Coordinating Conjunction)	<i>và, hoặc, hay, với (như hình với bóng, anh với tôi cùng đi, đến khổ với nó), ...</i>
16	C	SC	liên từ phụ thuộc (Subordinating Conjunction)	<i>như, nhưng, nếu, thì, vì, nên, ...</i>
17	T, Aux	PRT	trợ từ (Particle)	<i>à, ư, nhỉ, nhé, ...</i>
18	I	I	cảm từ (Interjection)	<i>ái_chà, ôi, chao_ôi, trời_ơi, haizzz, hihihi, hehe, ...</i>
19	G, X	ADJ	tổ hợp có tính chất như tính từ	<i>Bao gồm cả các thành ngữ có yếu tố không có nghĩa hoặc mơ hồ về nghĩa (như thập_tử_nhất_sinh, đầu_tất_mặt_tối,...):</i> <i>một_khác, được_chăng_hay_chớ, thập_tử_nhất_sinh...</i>
20	G, X	ADV	tổ hợp có tính chất như phụ từ, bổ nghĩa cho động từ, tính từ	<i>làm_gì, có_vẻ, ít_nhất, chỉ_ít, may_ra, hoạ_chăng, hoạ_may, một_mình, ngày_một, ngày_càng, càng_ngày_càng, đến_nổi, ai_ngờ, (học_tập) làm_sao (cho_tốt), ra_bộ, ra_cái_vẻ, ra_cái_điều, ra_điều, được_cái, ...</i>
21	G, X	N	tổ hợp có tính chất như danh từ	<i>địa_linh_nhân_kiệt, vừa_qua, vừa_rồi, mới_đây, mới_rồi, rồi_đây, rồi_mai, rồi_ra, một_mặt, mặt_khác, ...</i>

22	G, X	PRT	tổ hợp có thính chất như trợ từ	<i>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_chẳng_nữa</i>
23	G, X	SC	tổ hợp có thính chất như liên từ phụ thuộc	<i>làm_sao, may_mà, chẳng_hạn, khác_nào, chẳng_qua, đã_đành, đành_rằng, phải_chẳng, khôn_nỗi, khổ_nỗi, khổ_một_nỗi, khôn_một_nỗi, ...</i>
24	G, X	V	tổ hợp có thính chất như động từ	<i>tóm_lại, nói_tóm_lại, nói_riêng, nói_chung, nhìn_chung, coi_là, xem_là, kể_cả, thay_vì, (không ai) làm_sao, ...</i>
25	MW	MWE	ngữ cố định/thành ngữ (Multi-word expression)	Gán nhãn thành phần tùy 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))
26	D	D	tổ hợp từ biểu thị ý nghĩa tục tĩu (Dirty word)	- <i>dù_má, dù_mẹ, vãi_chuồng, đéch, đéo, vãi, ...</i>
27	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: <i>Cử đầu vọng minh nguyệt, Đê đầu tư cổ hương</i> (Ngẩng đầu nhìn trăng sáng, Cúi đầu nhớ cổ hương)
28	Z	Z	yếu tố cấu tạo từ (Bound morphemes)	- Các yếu tố đứng độc lập, không tham gia cấu tạo từ: <i>bắt, vô, phi, tổng...</i> (theo đó <i>bắt, phi, tổng</i> trong <i>bắt_hợp_tác, phi_nhân_đạo, tổng_kiểm_tra</i> là tham gia cấu tạo từ, không gán nhãn Z riêng cho chúng)
29	y	y	Từ viết tắt (Abbreviation, Acronym)	Viết tắt tên tổ chức, thuật ngữ <i>OPEC, WTO, HIV, AK, Ca, 3D, Mig-27, (mạng) 3G, GS., TS., tr., NQ-TW, ...</i> <i>v.v., T/L, T/M, ...</i>
30	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). <i>Internet/Nb, email/Nb, video/Nb, chat/Vb, sexy/ADJb, ...</i> - 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...
31	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. - <i>made, in, Thailand (made in Thailand), anyway, and, or, but, I, love, you, too (I love you too), how, are, you (how are you), ...</i>
32	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): ., ; : ? ! ... - () “” < > [] { }
33	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: <i>I., A., a), ii., v.v.</i> - Các dấu dùng trong công thức toán: +, -, ×, ÷, =, <, > - Các tổ hợp khác như công thức toán học, hoá học: $x^2 = a$, $x+y = 36$, $\sqrt{A^2} = A$, $AI \times \sqrt{A^2}$ - 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