Comprehensive Multi Level POS Tag set for Sinhala

This document presents a comprehensive, multi-level POS tag set for Sinhala. Examples are given for each tag.

1. Level 1

Sinhala words are broadly categorized in to five parts of speech at level 1

1. Nouns

2. Verbs

3. Adjectives

4. Adverbs

5. Nipāta

2. Level 2

At level 2, each tag at level 1 is further categorized based on context definitions.

2.1 Categories of Nouns

2.1.1 Common Nouns (NNC)

Common Noun denotes a class of objects or a concept. At second level of tag set, both animate and inanimate common nouns and their inflections belongs to the same category.

Examples: මිනිසා, ලේකම්, පාසැල, ආයතනය, සභාව, දැනුම, පදනම, මිනිසෙක්, මිනිසෙකු, මිනිසාගේ, මිනිසාට, මිනිසාගෙන්, මිනිසුනේ, පාසලට, පාසලෙහි, පාසලෙන්, පාසලක්, එරට, මෙරට.
Example usage: ප්‍රාදේශීය **සභාවේ** **ලේකම්**, පාසැල් **ගොඩනැගිල්ල** පරීක්ෂා කරයි.

2.1.2 Pronoun (PRP)

Pronouns are words that can be substituted for a noun or a noun phrase.

Examples: මම, අපි, ඔහු, ඇය, ඈ, තමා, ඌ, ඔවුන්, මට, මගේ, මගෙන්, එය, මෙය, මේවා, මෙයට, මෙයින්, මෙහි, දෙය, දෙයක්
Example Usage: **ඔහු** නව පාසල් ගොඩනැගිල්ල විවෘත කළේය. **එය** ඉතා අලංකාර ගොඩනැගිල්ලක් වන අතර **එයින්** පාසලට නව පෙනුමක් ලැබී ඇත.

2.1.3 Proper Noun (NNP)

Proper noun identifies an exact entity (person, place or thing).Since it refers to an exact known entity, proper nouns cannot have an indefinite form.

Examples: නිමල්, විමලා, ලංකාව, මොරටුව, ඇමරිකා එක්සත් ජනපදය, රුපියල, සිංහල, නිමල්ට, නිමල්ගේ, නිමල්ගෙන්, රුපියල්, රුපියලට, රුපියලෙන්

TIP: All parts of a Compound Proper Noun should be tagged as “Proper Noun (NNP)” except nipātha

- ○ ශ්‍රී_NNP ලංකා_NNP ප්‍රජාතාන්ත්‍රික_NNP සමාජවාදී_NNP ජනරජය_NNP
- ○ ශ්‍රී_NNP ලංකාව_NNP ප්‍රජාතාන්ත්‍රික_JJ රටකි_NNV

2.1.4 Questioning Pronouns (QUE)

Questioning Pronouns are words used to ask a question.

Examples: කුමක්ද, කෙසේද, කවදාද, කවරක්ද, කවරෙක්ද,

2.1.5 Deterministic Pronoun (NDT)

Deterministic pronouns are words built up from a combination of a determiner (discussed below) and a pronoun. For example සමහරෙක් (some of them) is a word in Sinhala derived from සමහර (some), which is a determiner and ඔවුන් (them), which is a pronoun.

Examples: ඇතැමෙක් (ඇතැම්+අයෙක්), සමහරෙක් (සමහර+දෙනෙක්), කිසිවෙක් (කිසි+කෙනෙක්), සමහරක් (සමහර+දෙයක්), කිසිවක් (කිසි+දෙයක්)

2.1.6 Question Based Pronoun(QBE)

Question base pronouns are used to show the uncertainty of a noun/noun phrase of interest.

Examples: කිම, කිමෙක්, කුමන, කුමක්, කුමකට, කුමකින්, කුමක, කවර, කවරක්, කවරකට, කවරකින්, කවරක, කවරෙක්, කවරෙකුට, කවරෙකුගෙන්, කවරෙකුගේ,

2 Categories of Verbs

2.2.1 Verb Finite (VFM)

Finite Verbs are verbs used at the end of a sentence.

Examples: බලමි, බලමු, බලති, බලන්නේය, බලන්නීය, බැලුවේය, බැලූහ, බැලිණ, බලන්න, බලනු.

2.2.2 Verb Participle (VP)

An inflected form of a verb that is used in a sentence to modify a noun, noun phrase, verb or a verb phrase is called a verb participle. Thus it plays the same role as adjective or adverb.

Examples: බැලූ, බලන, බැලුවේ, බැලිය, බැලෙන, බලන්නේ, බැලීමේ, බලනු
Example Usage:
සතියක් තුල අත්හදා **බැලූ** දෙවැනි මිසයිලය මෙයයි
එබී **බලන** මිනිසා
සොයා **බැලිය** යුතු බව

2.2.3 Verbal Noun (VNN)

An inflected form of a base verb that acts as a noun is a verbal noun.

Examples: බැලීමට, බැලීමටත්, බැලීමේ, බැලීමෙන්, බැලීම, බැලීමට, බැලීමක්, බැලුම්, බැලිලි

Example Usage: සිය මව **බැලීමට** බුරුමයට පැමිණි විට

පස් මහා **බැලුම්** බල

2.2.3 Modal Auxiliary (AU)

The Verbs which expresses necessity or possibility are called model auxiliary.

Example: හැකි, යුතු, නොහැකි

2.2.4 Verb Non Finite (VNF)

Verbs that does not belong to any of the above categories are considered as verb non finite

Examples: බලා, බලමින්, බලද්දී, බලතොත්, බැලුවාම, බැලුවත්
Example Usage:
තම නිවෙස් **බලා** යාමට
පිට වීමට මාන **බලමින්** සිටි

2.2.5 Compound Verbs

There are some **Verbs** in Sinhala which are **compounds of two words**. In such situations, second word when taken alone is always a Verb whereas first word can be a Noun, Adjective or a word which does not belong to any other form.

The first words of such compounds are categorized into four tags as below.

2.2.5.1 Noun in Compound Verbs (NCV)

If a compound verb is a combination of Noun + Verb, such nouns are identified as Noun in Compound verbs. Noun in compound verb is always the base, plural form of the particular noun. There is no corresponding translation in English, since all compound verbs in Sinhala is a normal verb in English Examples of compound verbs. First word of following compound verbs are identified as 'Noun in Compound Verb'

Examples: පාඩම් කරනවා
සෙල්ලම් කරනවා
වර්ග කරනවා
ප්‍රතිඥා දෙනවා
බලාපොරොත්තු වෙනවා

2.2.5.2 Adjective in Compound Verbs (ACV)

If a compound verb is a combination of Adjective + Verb, such adjectives are identified as Adjective in Compound verbs. First word in such compound verbs will be tagged as Adjective in compound verbs

Examples: අඩු කරනවා/ කළ/ වූ
වැඩි කරනවා
සමාන කරනවා
ක්‍රියාකාරී වෙනවා
අඩපන වී
අසාර්ථක වීමේ

2.2.5.3 Particle in Compound Verbs (RPCV)

Preposition in compound verb is a word that doesn't have a meaning by itself but, when combined with another verb, make up a compound verb.

Examples: ඉටු කරනවා/වෙනවා/ කළ/ වූ
සිදු කරනවා

TIP: **Noun/Adjective/ Particle in Compound Verb** is always followed by a Verb type (VFM, VNF, VP or VNN)

2.2.5.4 Supportive Verb in Compound Verb (SVCV)

Some compound verbs are consisted with a main verb and a supportive verb. Specially, ගනියි/ගන්නවා / තිබේ / තියෙනවා act as supportive verbs in compound verbs to generate a one meaning. In such cases, the second verb is tagged as SVCV.

Examples: බලා **ගනියි** , දැක **තිබේ**

2.3 Categories of Adjectives

2.3.1 Adjective (JJ)

Adjectives are words used to describe nouns or noun phrases. Whole purpose of an adjective is to describe a noun and could not be used as any other word form.

Examples: රළු, සුමුදු, වැඩි, හොඳ, විශේෂ, අලුත්, සුදු, අඩු, විශාල, කුඩා, ඉහළ, දිග, පුංචි, ලොකු, අධික, දියුණු, සුළු, යට, සුදුසු, උඩ, මහා, උසස්, ප්‍රධාන, නියම, මුලික, තරුණ, පහළ, ප්‍රකට, මහත්, පැරණි, නිසි, මානව, විවෘත

- පළමු, දෙවන, තෙවන
- ප්‍රාදේශීය, නාගරික, ළමා, අධ්‍යාපන

2.3.2 Adjectival Noun (NNJ)

Adjectival noun is a common noun that acts as an adjective to describe another noun. When a common noun is used as an adjectival noun, it always takes the base, plural form of the common noun. So in a noun phrase like ‘පාසල් වත්ත’, ‘පාසල්’ is an adjectival noun which describes the main common noun ‘වත්ත’

Examples with usage: **චිත්‍ර** පොත, **ගල්** කණු, **පාසල්** වේලාව,
Example Usage: ප්‍රාදේශීය සභාවේ ලේකම්, **පාසැල්** ගොඩනැගිල්ල පරීක්ෂා කරයි.

TIP: Adjectival Nouns always take the base form of the noun. In Sinhala, base form of all

TIP: **Adjective or Adjectival Noun??**

Adjectives (JJ) can only be used as adjectives regardless of context. Whereas Adjectival Nouns (NNJ) can be used as nouns (NNP, NNC) alone, without any change to the word.

Ex: **Adjectival Noun**

පාසැල් can be an Adjectival Noun or Common Noun.

මම **චිත්‍ර_NNJ** පොතක **චිත්‍ර_NNC**ඇන්දෙමි.

2.4 Categories of Adverbs (RB)

Adverbs are not further sub categorized in second level.

Examples: වේගයෙන්, සෙමින්, උද්යෝගයෙන්, එකට, එසේ, මෙසේ, කෙසේ, එබැවින්, දැන්, තවමත්, නැවත, නැවතත්, කලින්, එහෙයින්, එසේම, මෙතෙක්, එහෙම, විශේෂයෙන්, යලි, දැනට, වහා, දැනටමත්

2.5 Categories of Nipāta

2.5.1 Postposition (POST)

Postpositions in Sinhala are words used after nouns, verbs and sometimes even after adjectives and adverbs to show their relationship to other words in order to build up a meaningful sentence.

Examples: සඳහා, අතර, ලෙස, සිට, වැනි, තුළ, විසින්, අනුව, ගැන, මෙන්, මත,පිළිබඳව, වෙත, සේ, මිස, උදෙසා, සම්බන්ධයෙන්, මගින්, දක්වා, තුළින්, පිණිස, යටතේ, පිට, වෙනුවෙන්, අතරින්, ඔස්සේ, පටන්, හරහා, හට, හැර, මෙන්ම, වෙනුවෙන්, වෙනුවට, තෙක්, පරිදි, හෙයින්, හෙවත්, බැහැර, හැටියට, පුරා, කරා, සමග, එක්ක, ආදී,පමණක්, සහිත, රහිත , සහිතව, රහිතව Number + වන/වැනි/වෙනි.

2.5.2 Conjunction (CC)

Conjunctions are words used to connect words, phrases or sentences.

Examples: හා, හෝ, නමුත්, යැයි, යයි, දැයි, එහෙත්, ඒත්.

2.5.3 Particle (RP)

Particle is a word or a part of a word that has a grammatical purpose but often has little or no meaning.

Examples: ය, ම, ද, දී, යි, ලු, ඉතා, වඩා, වඩාත්, ක

2.5.4 Interjection (UH)

Interjections are words used to show the emotion or feeling.

Examples: අහෝ, චිහ්, ෂික්, අහ්

2.5.5 Nipathana (NIP)

Nipathana can be used alone in some contexts, and can be used as a postposition as well if needed.

Examples: ඇති, නැති, ඕනෑ, එපා, බැරි, පුළුවන්, ඇතිව, නැතිව, ඇත්නම්.

2.5.6 Determiner (DET)

Determiners are words that are used before a noun to show which particular example of the noun is referred to.

Examples: මේ, ඒ, මෙම, බොහෝ, එක්, සමහර, ඔය, යම්, අර, ඇතැම්, ටික, සියලුම, මුළු, කවර, හැම, අන්, කිසි, තව, වෙන, කිසිම, වෙනත්, අනෙක්, කිසිඳු, මුළු, සෑම, එම, කිසියම්, අනෙක්,
Example Usage: **මෙම** පාසැල ගමේ **අනෙක්** පාසල් අතර කැපී පෙනෙයි.
ඇතැම් දරුවන් ක්‍රීඩා කරති
අද උත්සවයක් පැවැත්විණි. **කිසිඳු** කෙනක් **ඒ** උත්සවය සඳහා පැමිණියේ නැත

2.5.6 Case Marker (CM)

Sinhala nouns are morphologically inflected based on the case. A suffix is added to the noun to show the case. For animate nouns and inanimate singular nouns, suffix ට - ta is added for dative case, suffix ගේ - "ge" is added for genitive case and suffix ගෙන් - "gen" is added for instrumental case. For inanimate plural nouns, suffix වලට - "valata" is added for dative case, suffix වල - "vala" is added for genitive case and suffix වලින් - "walin" is added for instrumental case. According to Sinhala language rules, it is wrong to separate these case marking suffixes from the main noun. However, some Sinhala writers tend to separate this case marking suffix from the main noun. If separated, such suffixes are categorized under case markers. Case marker does not have an English meaning on its own.

නිමල්ට/ නිමල් ට , නිමල්ගේ/ නිමල් ගේ, පොත්වලට/ පොත් වලට
Since we need to handle both of these inputs, a POS tag called 'Case Marker' is introduces to tag the suffix, if written separately.
Example: ට, ගේ, ගෙන්, වලට, වලින්, වල, හිදී

2.6 Extra tags at Level 2

2.6.1 Sentence ending (NNV)

A word (Noun, Postposition, Adjective, etc.) that is used to end the sentence, but it is not a finite verb. When such a word is used as a sentence ending, original word is inflected with prefixing a particle like යි (yi) or ය (ya).

Examples: පොතකි, ගසකි, ශාලාවකි, ගෙයකි,
Example Usage: ඔහු කියවා ඇත්තේ **පොතකි**.

2.6.2 Number (NUM)

All numbers written in numeric or letters

Example: එක, දෙක, 128, 1.28, එක්

2.6.3 Abbreviation (ABB)

Example: අ පො ස

2.6.4 Full Stop (FS)

2.6.5 Punctuation (PUNC)

TIP: **Full Stop or Punctuation?**

Tag "Full stop / Period" is to mark the end of sentence. When "." Is used for other purposes (Ex: 22.34, අ. පො. ස., ඒ. බි. සී. පෙරේරා), it should be tagged as punctuation

2.6.6 Foreign Word (FRW)

A word that is written in any other language .

Examples: BOC , ISBN , NFC , Japan media communication

2.6.6 Tag Undefined (TAG UNDEFINED)

The words which cannot be tagged under any of the above categories are tagged under this tag.

3. Level 3

Inflection based grammatical variations of the language is captured using tags at third level. Sinhala words (Nouns and verbs) can be inflected based on Animacy, Number, Gender, Person, Case, Tense and Definiteness.

3.1 Common Noun Inflections

Common nouns can be either Animate or Inanimate. Those two categories are further sub categorized based on other inflecting factors. Each cell of the following tables is a separate tag.

Animate common nouns can be inflected based on number, gender, definiteness and case (Nominative, Accusative, Dative, Genitive, Instrumental).

3.2 Categories of Nipāta

3.2.1 Particle

3.2.1.1 Verbal Suffix (VSX)

In some cases, the writers attend to separate the two particles ය,යි in the place of verbs instead of verbs. Such particles act as the verbs. They are called as Verbal Suffixes.

Examples: ය ,යි

3.2.1.2 Adverbial Suffix (AVS)

The particle ව is attached to the adjectives as well as adverbs to modify the meaning further. In some cases, the writers tend to write the particle separately as well as attached to the previous word. This tag is used to define the article where it is written separately from the previous word.

Examples :ව

3.2.1.3 Negative Prefix (NGP)

The prefix නො is attached to the verbs,participles as well as adverbsnouns to modify the meaning further. In some cases, the writers tend to write the prefix separately as well as attached to the previous word. This tag is used to define the prefix where it is written separately from the previous word.

Examples: නො කරයි , නො සිතූ

3.2.1.4 Prefix (PRF)

Sinhala language is consisted with main prefixes (upasarga) as ප, පර, අව, ස, අනු, නි, දු, වි, උප, අප, පස්, පිළි, පි, අති, අදි, අබි. And some writers tend to write the prefix separately as well as attached to the previous word. This tag is used to define the prefix where it is written separately from the previous word.

Examples: අති ක්‍රමණ

3.2.1.5 Particle in Quotation (RPQ)

This particle is used in the middle of a sentence to connect two sentences.

Examples: **(…………..’යි……**)