TAMIL PARTS-OF-SPEECH TAGGING GUIDELINES

Designing Tagsets and Specifications

Part of speech (POS) tagging is the process of assigning a unique part of speech to each word (token) in a sentence. This process helps in identifying and disambiguating the role of each word (token) in a sentence. POS tagger tags words in a context using POS tags.

This POS guideline describes the different types of tags that are needed to tag the Tamil data set. The rules developed for each tag are language specific. This set of POS guidelines was framed based on the morphological, syntactic and semantic structure of the sentence. It has described a set of 16 tags with examples as seen below:

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

1. **ADJ: adjective**

Adjectives are words that modify nouns. The following are various occurrences of ADJ in Tamil. The word ends with -**ஆன** are tagged adjectives, describing various features as follows:

(i) <u>Attributive adjective</u>: Attribute adjectives provide additional information about a noun by describing its characteristics. It occurs before the noun.

quality (அழகான/கூர்மையான),
size/quantity (பெரிய/ சிறிய;கொஞ்சம்/நிறைய),
shapes (வட்ட/சதுர),
age/time (பழைய/புதிய;வார/மாத),
material (தங்க/வெள்ளி),
emotions (கோபமான/மகிழ்ச்சியான),
position (முதலாம்/நூறாவது)

(a) Quality

Example:
அது உயரமான மரம்.
அவள் சுறுசுறுப்பான மாணவி.

(b) Size/quantity

Example:
இங்கு நிறைய மரங்கள் உள்ளன.
இது மிகவும் பெரிய கோவில்.

(c) Shapes /colour

Example:
பழங்கள் உருண்டை, நீள்வட்டம், வட்ட வடிவங்களில் காணப்படும்.
சிவப்பு ரோஜா அழகாக உள்ளது.

(d) Age/time

Example:
கோட்டையில் பழைய பிரிட்டிஷ் இராணுவ நிர்வாகக் கட்டிடங்கள் உள்ளன.

(e) Emotions

Example:
வெற்றி பெறுவது மிகவும் மகிழ்ச்சியான தருணம் ஆகும்.

(f) Position

Example:

சீதா வகுப்பில் மூன்றாவது உயரமான பெண்.

(ii) <u>Numbers vs. Adjectives</u>: In general, cardinal numbers receive the part of speech NUM, while ordinal numbers (more precisely adjectival ordinal numerals) receive the tag ADJ.

Example:

நான் முதல் மதிப்பெண் பெற்றேன்.

(iii) <u>Adjectival modifiers of adjectives</u>: In general, an ADJ is modified by an ADV (மிகுந்த பலசாலி). However, sometimes a word modifying an ADJ is still regarded as an ADJ. These cases include:

(a) ordinal numeral modifiers of an adjective :

Example:

மெரினா கடற்கரை உலகின் இரண்டாவது நீளமான கடற்கரை.

(b) when a pair of adjectives/number+adjectives form a compound adjectival modifier :

Example:

சிட்டுக்குருவி ஒரு சுறுசுறுப்பான அழகான பறவை ஆகும்.

Syntactic cue:

It is followed by another adjective or a noun/pronoun

All **-ஆன** ending words and not **-க்கான** (as it is considered as NOUN)

Test case:

அவன், அவள், அது can be added to the given word if the word is adjective

2. ADP: adposition

Adposition is a cover term for prepositions and postpositions. Adpositions belong to a closed set of items that occur before (preposition) or after (postposition) a complement composed of a noun phrase, noun, pronoun, or clause that functions as a noun phrase, and that form a single structure with the complement to express its grammatical and semantic relation to another unit within a clause.

Example:
பார்வையாளர்கள் காலை 8.00 மணி முதல் மாலை 6.00 மணி வரை கோட்டைக்குள் நுழையலாம்.

List of ADPs: அடியில், அப்பால், அருகிலேயே, அருகே, ஆகச், இடையில், இடையே, இன்றி, இருந்து, இருந்தே, உட்பட, உட்பட்ட, உள், உள்ள, உள்ளே, எதிரெதிரே, எதிரே, ஏற்ப, ஒட்டி, கீழே, கீழ், கீழ்க், குறித்து, குறுக்கே, கூடவே, கொண்ட, கொண்டு, சுற்றி, சுற்றியும், சுற்றியுள்ள, தவிர, நடுவில், நடுவே, பற்றி, பற்றிய, பற்றியும், பின்னால், பிறகு, போது, போதும், மத்தியில், மீது, முதல், முன், முன்னால், முன்பு, மூலமோ, மூலம், மேலே, மேல், வரை, வரையிலும், விட, etc.

Syntactic cue:
It is preceded by NOUN (defining the position of the noun)

3. **ADV: adverb**
Adverbs are words that typically modify verbs for such categories as time, place or manner. In Tamil, the word ending with **-ஆக** is an adverbial cue. They may also modify adjectives and other adverbs, as in **மிகப் பெரிய** or **மிகவும் வேகமாக**. But **-க்காக** endings are NOUN and not ADV, as in **அமைதிக்காக**.

Types:

(i) Interrogative/relative adverbs: எங்கே, எப்பொழுது, எப்படி, ஏன், எவ்வாறு (including when marks a clause that is circumstantial, not interrogative or relative)

Example:
குமார் ஏன் இன்னும் வரவில்லை?

(ii) Demonstrative adverbs: இங்கே, அங்கே, இப்பொழுது, அப்பொழுது

Example:
நீ வீசிய பந்து இங்கே உள்ளது.

(iii) Totality adverbs: எங்கேயும், எப்பொழுதும்

Example:

அவன் எப்பொழுதும் தாமதமாக வருவான்.

(iv) Adverbs of manner:

Example:

பல நூற்றாண்டுகளாக மக்கள் ஆர்வமாக இங்கே வாழ்ந்து வந்துள்ளனர்.

(v) Adverbs of time:

Example:

அவன் தற்போது வீட்டில் உள்ளான்.

(vi) Adverbs of frequency:

Example:

நான் அடிக்கடி கண்ணாடியைப் பார்ப்பேன்.

(vii) Adverbs of degree:

Example:

அவன் மிக அழகாக வரைகிறான்.

(NOTE: all directions are tagged as NOUN which are usually adverbs)

Syntactic cue:

Words with -**ஆக** endings

Test case:
The word order in the sentence can be changed and the meaning remains the same.

4. **AUX: auxiliary**
An auxiliary is a function word that accompanies the lexical verb of a verb phrase and expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect, voice or evidentiality. It is often a verb (which may have non-auxiliary uses as well) but many languages have nonverbal TAM markers and these should also be tagged AUX. The class AUX also includes copulas (in the narrow sense of pure linking words for nonverbal predication).

Some modal verbs may count as auxiliaries in Tamil:
Example:
கோப்பையை பத்திரமாக வைத்து இருந்தனர்.

Tense auxiliaires:
Tense auxiliaries express when the action is taking place. It is especially seen while expressing continuous tense in Tamil.

Example:
அவன் வேகமாக வந்து கொண்டு இருந்தான்.

Note: In some cases, கொண்டு is the main verb (VERB) and not AUX.

Example:
சென்னை பல அழகிய கடற்கரைகளைக் கொண்டு உள்ளது.

Passive auxiliaries:
Passive auxiliaries are to be split for tagging since they occur together with the main verb in this data.

Example:
கடற்கரை ஆழமற்ற நீரால் ஆசீர்வதிக்கப்பட்டுள்ளது.

List of auxiliaries:
இரு, வேண்டு, கொள், விடு, மாட்டு, வா, வை, கூடு, படு, முடி, செல், செய், உள்ள, ஒடு, இழு, இல்லை, அல்ல, பார், etc.

Note: In cases like உட்பட்டது, வெளிப்படுத்தும், பயன்படுத்து, படு is not split separately. It is one whole word.

5. **CCONJ: coordinating conjunction**
A coordinating conjunction is a word that links words or larger constituents without syntactically subordinating one to the other and expresses a semantic relationship between them. In Tamil, அல்லது and மற்றும் are the CCONJs tagged.

Examples:
நான் மற்றும் என் நண்பன் மற்ற மாணவர்களுடன் செல்லவில்லை.
எனக்கு ஆப்பிள் அல்லது ஆரஞ்சு வேண்டும்.

6. **DET: determiner**
Determiners are words that modify nouns or noun phrases and express the reference of the noun phrase in context. That is, a determiner may indicate whether the noun is referring to a definite or indefinite element of a class, to a closer or more distant element,

to an element belonging to a specified person or thing, to a particular number or quantity, etc.

The following types are found in Tamil:

(i) Demonstrative determiners:

Example:
பூக்கள் அந்த இடத்தை மிகவும் அழகாக ஆக்குகிறது.

(ii) Interrogative determiners:

Example:
நீங்கள் எந்த நாட்டிற்குச் செல்கிறீர்கள்?

(iii) Quantity determiners:

Example:
சில நேரங்களில் நேரம் வீணாவதே தெரிவதில்லை.

List of DETs:
அக், அந்த, அனைத்து, அனைவருக்கும், அப் ,இக், இது, இத், இந்த, இப், இம், இரு, எந்த, எல்லாம், ஒரு, ஒவ்வொரு, சில, பல, பலவற்றின், பல்வேறு, பிற, முற்றிலும், முழு, வேறு, etc.

7. **INTJ: interjection**
An interjection is a word that is used most often as an exclamation or part of an exclamation. It typically expresses an emotional reaction, is not syntactically related to other accompanying expressions, and may include a combination of sounds not otherwise found in the language.

Example:
அப்பப்பா! அதன் சுவையே தனி தான்.

8. **NOUN: noun**
Nouns are a part of speech typically denoting a common thing, animal, plant or idea. In Tamil, gerunds share the quality of noun and it is tagged as NOUN.

Example:
பழங்களில் நிறைய வகைகள் உள்ளன.

Some of the special cases are considered as NOUN:

(i)Gerunds: All gerunds/ gerund+case marker or without case marker is considered as NOUN

Example:
மூச்சுப்பயிற்சி செய்தல் உடல்நலத்திற்கு நல்லது.

(ii) Pronominalised participle verbs:

Example: இருந்தவன், இருக்கிறவன், இருப்பவன், இருக்காதவன் followed by any case marker.

Example:

நேற்று படிக்காதவனை இன்று படிக்கச்சொல்.

(iii) Oblique forms:

In Tamil, the oblique forms of words are considered as noun rather than adjectives.

Example:

புனித நீர்

Here, the root word of **புனித** is **புனிதம்**. So, **புனித** is in oblique form. Such cases are considered as nouns and not adjectives.

(iv) Directions:
All directions are included in NOUNs as they belong to Nouns of space and time.
Example:
கிழக்கு தமிழகத்தில் இன்று மழை பெய்யும்.

(v) Predicative adjective:
The adjective acts like a noun at the end of a sentence in Tamil.

Example:
அந்த மருத்துவர் மிகவும் நல்லவர்.

NOTE: adverbs end with **-ஆக** but nouns end with **-க்காக**

Example: **அமைதியாக** is ADV and **அமைதிக்காக** is NOUN

9. **NUM: numeral**

A numeral is a word, functioning most typically as a determiner, adjective or pronoun, that expresses a number and a relation to the number, such as quantity, sequence, frequency or fraction.

Note that cardinal numerals are covered by NUM whether they are used as determiners or not (as in Windows Seven) and whether they are expressed as words (four), digits (4) or Roman numerals (IV)

NUM includes numeric date/time formats (11:00) and phone numbers. Words mixing digits and alphabetic characters should, however, ordinarily be excluded. In English, for example, pluralized numbers (the 1970s, the seventies) are treated as plural NOUNs, while mixed alphanumeric street addresses (221B) and product names (130XE) are PROPN. But in Tamil, since they are split, the numbers are tagged as NUM and followed by PART for the words like -ஆம் and -கள் .

Example:
மதுரை நகரத்திலிருந்து 30 கிமீ தொலைவில் உள்ள இந்த இடத்தை வந்தடைவதற்கான சாலைகள் நன்கு இணைக்கப்பட்டுள்ளது .

List of NUMs:
Numbers/digits: **0, 1, 2, 3, 4, 5, 2014, 1000000, 3.14159265359**

Date/time: **11/11/1918, 11:00**

Word forms: ஒன்று, இரண்டு, மூன்று,எழுப்பது ஏழு

Tamil numerals: **க (1)**

Roman numerals: **I, II, III, IV, V, MMXIV**

10. **PART: particle**

Particles are function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech (e.g. adpositions, coordinating conjunctions, subordinating conjunctions or auxiliary verbs). Particles may encode grammatical categories such as negation, mood, tense etc.

Particles are normally not inflected, although exceptions may occur. In general, the PART tag should be used restrictively and only when no other tag is possible.

In Tamil, it is generally tagged after a participle form of a verb.

Example:

நான் வரும் போது அவன் தூங்கினான்.

அவன் இளமையாகவும் அழகாகவும் இருக்கிறான்.

List of PARTs: ஆம், இல், உம், களின், களில், கூடிய, க்கும், தான், பிறகு, போது, etc.

11. PRON: pronoun

Pronouns are words that substitute for nouns or noun phrases, whose meaning is recoverable from the linguistic or extralinguistic context.

(i) In Tamil, non-possessive personal, reflexive or reciprocal pronouns are always tagged PRON.

(ii) Possessives vary across languages. In Tamil, they are more like a normal personal pronoun in a specific case (often the genitive), or a personal pronoun with an adposition; they are tagged PRON.

(i) personal pronouns: நான், நாம், நீ, அவர்கள், அவன், அவள்

Example:

அவர்கள் நாணயங்களை அச்சிட்டனர்.

(ii) reflexive pronouns: தனக்குத்தானே

Example:

அவள் தனக்குத்தானே சிரித்துக்கொண்டாள்.

(iii) interrogative pronouns: யார்

Example:

மழையை விரும்பாதவர் யார் உள்ளார்?

(iv) possessive pronouns: என்னுடைய, நம்முடைய, அவர்களுடைய, உன்னுடைய

Example:

இது என்னுடைய பையில் இருந்தது.

(v) attributive possessive pronouns: எனது/என், உனது/உன்

Example:

உன் கண்களை மூடு.

12. PROPN: proper noun

A proper noun is a noun (or nominal content word) that is the name (or part of the name) of a specific individual, place, or object.

Example:
தமிழ்நாட்டின் புகழ்பெற்ற சுற்றுலாத் தலங்களில் மிகவும் பிரபலமான ஒன்றானது கொடைக்கானல்.

- Multi word names like ஐக்கிய அரபு நாடுகள் are tagged as PROPN for the specific word and நாடுகள் is tagged as NOUN

  Example:
  பிரபலமான மலை நகரமான பெல்லிகல் அருவிக்கு அருகில் அமைந்துள்ளது .
  Here, பெல்லிகல் is PROPN and அருவிக்கு is NOUN

- Acronyms of proper nouns, such as ஐநா and யுனெஸ்கோ, should be tagged PROPN.

13. PUNCT: punctuation

Punctuation marks are non-alphabetical characters and character groups used in many languages to delimit linguistic units in printed text.

Punctuation is not taken to include logograms such as *$*, *%*, and *§*, which are instead tagged as SYM. (Hint: if it corresponds to a word that you pronounce, such as *dollar* or *percent*, it is SYM and not PUNCT.)

Spoken corpora contains symbols representing pauses, laughter and other sounds; we treat them as punctuation, too. In these cases it is even not required that all characters of the token are non-alphabetical. One can represent a pause using a special character such as #, or using some more descriptive coding such as *[:pause]*.

Example: **.,()„„”,;,:,?,!**

14. SCONJ: subordinating conjunction

A subordinating conjunction is a conjunction that links constructions by making one of them a constituent of the other. The subordinating conjunction typically marks the incorporated constituent which has the status of a (subordinate) clause.

(i) Complementizers:

Example:
ஒகேனகல் என்பதற்கு கன்னட மொழியில் புகைப் பாறை என்று பொருள்.

(ii) Simultaneous construction:

Example:
கணேஷ் படித்துக்கொண்டிருந்த பொழுது/போது நான் சமைத்தேன்.

(iii) Discourse Connector:

Example:
எனவே, செயின்ட் ஜார்ஜ் கோட்டை என்று பெயரிடப்பட்டது.

List of SCONJs: அதனால், அதற்கு, ஆனால், இதனால், இருந்தபோதிலும், இருப்பினும், என, எனவே, எனும், என்பதற்கு, என்பதால், என்பதில், என்பது, என்பதே, என்பதை, என்ற, என்றால், என்று, என்றும், ஏனெனில், ஏனென்றால், காட்டிலும், தவி, பின்னர், போன்ற, போல, போலவே, போல், மேலும், etc.

15. SYM: symbol

A symbol is a word-like entity that differs from ordinary words by form, function, or both.

Many symbols are or contain special non-alphanumeric characters, similarly to punctuation. **What makes them different from punctuation is that they can be substituted by normal words.** This involves all currency symbols, e.g. *$ 75* is identical to *seventy-five dollars*.

Mathematical operators form another group of symbols.

Another group of symbols is emoticons and emoji.

Characters used as bullets in itemized lists *(•, ‣)* are not symbols, they are punctuation.

Examples:

$, %, §, ©

+, −, ×, ÷, =, <, >

:), ♥‿♥, 😝

john.doe@universal.org, http://universaldependencies.org/, 1-800-COMPANY

16. VERB: verb

A verb is a member of the syntactic class of words that typically signal events and actions, can constitute a minimal predicate in a clause, and govern the number and types of other constituents which may occur in the clause.

Verbs are often associated with grammatical categories like tense, mood, aspect and voice, which can either be expressed inflectionally or using auxiliary verbs or particles.

Example:

நான் என் வீட்டுப்பாடத்தைச் செய்தேன்.

Note that the VERB tag covers main verbs *(content verbs)* but it does not cover *auxiliary verbs* and verbal *copulas* (in the narrow sense), for which there is the AUX tag. *Modal verbs* are considered AUX in Tamil.

Participles are word forms that may share properties and usage of adjectives and verbs. In Tamil, such constructions are considered VERB. Similarly, infinitives are also considered as VERB.

Participle construction:

Example:

[[(உங்கள் காலை உணவை) (உங்கள் நாவிற்கு) (விருந்தளிக்கும் விதத்தில்) சாப்பிட்டு] மகிழலாம்.]

Infinitive construction:

Example:

மேலும் அந்த இடத்தை தனுஷ்கோடியிலிருந்து பார்க்க முடியும்.