# **Annotation guidelines**

## **Dataset summary:**

The name of the dataset is ‘annotation\_tamil\_psych\_verbs’. It is in excel format (.xlsx). This dataset is a collection of annotations of Tamil sentences containing the psych verbs. The main purpose of creating the dataset is to analyze the word order pattern of Tamil psych verbs and to determine which pattern is most prevalent, whether it is SOV or OSV.

The dataset contains 1412 rows and 19 columns. Rows represents the data and columns represents the annotation patterns. These annotation patterns are the main factors affecting the word order.

## **Annotation:**

Below are the column names and their annotation guidelines.

1. **ID:**

This column contains the unique ID, which identifies the sentence. The values of the column are finite integers that starts from 1 to 1412.

1. **Verb:**

This column contains the lemma of psych verbs. There are 10 psych verbs in this column and they are ‘விய (viya - to wonder), மகிழ் (magizh - to be happy), விரும்பு (virumbu - to like), களி (kali - to ecstasize), நெகிழ் (negizh - to delight), அஞ்சு (anju - to fear), திகை (thigai - to shock), தவி (thavi - to suffer), வருந்து (varundhu - to be sad) and வெறு (veru - to hate)’.

1. **Sentence:**

This column contains the sentences in Tamil. (1) is the sample sentence from the dataset.

1. அங்கே அவர்கள் லோகாம்பிகையைக் கண்டு மகிழ்ந்தனர்.
2. **EO / ES:**

This column defines the type of psych verbs. Psych verbs are classified into 2 major groups such as EO and ES. The values of the column are ‘EO’ and ‘ES’. EO means Experiencer Object and ES means Experiencer Subject. When the object of a sentence is an experiencer, then the value is EO. When the subject of a sentence is an experiencer, then the value is ‘ES’.

In (2), object (Chidhambaram) is an experiencer and hence the value of this column is ‘EO’.

1. சிவந்தியின் கோபம் **சிதம்பரத்தை** வெகுவாக

*Sivanthiyin kobam* ***Chidhambarathai*** *veguvaaga*

of.Sivanthi anger (S) **Chidhambaram**(O) fastly

அச்சுறுத்தியது.

*achurithiyadhu.*

frightened (V).

The anger of Sivanthi frightened Chidhambaram fastly.

1. **Word\_order:**

This column contains the word order pattern of psych verb sentences. The values are ‘SOV’ and ‘OSV’. SOV means Subject Object Verb and OSV means Object Subject Verb. When the sentence is having SOV order, then the value is ‘SOV’. Else, it will be ‘OSV’.

In (3), the order of sentence is OSV and hence the value of this column is ‘OSV’.

1. அணில் வந்த\_காரணத்தை அறிய அனைவரும் விரும்பினர்.

*anil vandha\_kaaranathai ariya* *anaivarum virumbinar.*

squirrel reason.of. arrival to.know(O) everyone (S) wanted (V).

Everyone wanted to know the reason of squirrel’s arrival.

1. **Phrasal\_subject:**

This column indicates whether the subject is phrasal or not. The values of the column are ‘yes’ and ‘no’. In a sentence, the subject can be a single-word noun: அரசர் (*arasar*-King) or pronoun: அவன் (*avan* - He) or phrasal noun: (4) is the sample phrasal noun.

When the sentence contains phrasal subject, then the value is ‘yes’. Otherwise, it will be ‘no’.

In Tamil, the place and time occurs before the subject predominantly. Subject with time: சங்க காலத்தில் மக்கள் (*sanga\_kaalathil makkal -* people in sangam period) and subject with location: மெரினாவில் பல்லாயிரக் கணக்கில் மக்கள் (*merinavil pallaayira kanakkil makkal* – thousands of people in Merina beach). These kinds of subjects are also considered as phrasal.

1. காட்டு\_விலங்குகளை வேட்டையாடச் சென்ற ஒரு சோழ\_அரசன்

*kaatu\_vilangugalai vaetaiyaada sendra oru chozha\_arasan*

wild.animals to.hunt went one chola.king

One chola king who went to hunt the wild animals.

1. Clausal\_object:

This column indicates whether the object is clausal or not. The values of this column are ‘yes’ and ‘no’. In general, an object can be a single-word noun: பெண்ணை (*pennai -* women), pronoun: என்னை (*ennai -* me), phrasal noun: அரசியல் தலைவர்களை (*arasiyal thalaivargalai -* government officers) or clausal noun. Clausal noun is the one which has a verb in it.

In (5), bold words are the clausal noun that is an object. It has its own verbal form ‘**கண்டு** (*kandu* – by seeing)’ and hence it is a clausal noun. So, the column’s value is ‘yes’.

If the object of a sentence is clausal noun, then the value is ‘yes’. The prevalence of clausal object is high in Tamil psych verb sentences. If the object is other than clausal i.e., single-word noun or pronoun or phrasal noun, then the value is ‘no’.

1. **உங்கள் திறமையை கண்டு** அடுத்தவர்கள் வியப்பார்கள்.

***ungal thiramaiyai kandu*** *aduthavargal viyapargal.*

**your talent by.seeing (O)** others (S) will.wonder (V).

Others will wonder **by seeing your talent**.

1. **Case\_of\_subject:**

This column contains the cases of subject. In general, subject is always a nominative case. The values of the column are ‘nominative’ and ‘nominative (phrasal)’. When the subject is not phrasal, this column value will be ‘nominative’. Else, the value will be ‘nominative (phrasal)’.

1. **Case\_of\_subject\_inTamil:**

This column contains the Tamil grammatical terms for subject’s case. In English, the case of subject is called nominative. In Tamil, it is called முதல் வேற்றுமை (*1st vaetrumai*) or எழுவாய் வேற்றுமை (*yezhuvai vaetrumai*). The values of the column are ‘1st vaetrumai (yezhuvai)’, ‘ennummai’ and ‘mutrummai’.

For ‘1st vaetrumai (yezhuvai)’, no particles are affixed to the noun. When the subject is

affixed with particle ‘உம் (*um*)’ like ‘அனைவரும் (*anaivarum - everyone*)’, ‘எல்லோரும் (*ellorum - all*)’, ‘பலரும் (*palarum* - many)’ etc., then the value is ‘mutrummai’. If the subject is more than one like ‘அவனும் அவளும் (*avanum avalum* – He and she)’ or in other words if the subject has more than one ‘உம் (*um*)’, then the value is ‘ennummai’.

1. **Case\_of\_object:**

This column contains the cases of object. There are totally 18 values in this column and the most frequent ones are ‘verbal participle clause’, ‘special verbal participle ‘endru’ and ‘accusative’.

When the object is not clausal, then the value can be ‘accusative’, ‘accusative (phrasal)’, ‘dative’, ‘dative (phrasal)’ or ‘adverbial phrase’. When the object is clausal, the value can be ‘verbal participle clause’, ‘special verbal participle’, ‘negative verbal participle clause’, ‘special negative verbal participle’, ‘infinite verbal clause’, ‘accusative participle’, ‘dative participle ’or ‘adverbial clause'. The special verbal participle is divided into 5 groups such as ‘endru’, ‘ena’, ‘aai’, ‘aaga’ and ‘patri’. The special negative verbal participle is divided into 2 groups such as ‘illamal’ and ‘indri’.

In Tamil, the accusative case has particle ‘ஐ (*i*)’ at the end. In (6), bold words are the object and it is the ‘accusative (phrasal)’ case. The dative case has particle ‘கு (*ku*)’ at the end. (7) is the dative case. The adverbial phrase is formed by affixing ‘ஆல் (*aal*)’, or ‘இல் (*il*)’ to the accusative.

This verbal participle is formed by affixing ‘உ (*oo*)’ to the non-finite verb. In (8), bold words are the verbal participle clause as it contains ‘இழந்து (*izhandhu* - having lost)’ verbal participle. The negative verbal participle is having ‘அல் (*al*)’ or ‘ஆது (*aadhu*)’ at the end of non-finite verb. In (9), bold words are the negative verbal participle clause.

When ‘ஆல் (*aal*)’, or ‘இல் (*il*)’ is affixed to the verbal participle, it becomes adverbial clause - reason. When ‘போது (*podhu*)’, ‘பொழுது (*pozhudhu*)’, ‘உம் (*um*)’ is affixed, it becomes adverbial clause - timing. In some cases, ‘இல் (*il*)’ also defines the timing. In (10), the bold words are the adverbial clause (reason).

The infinite verbal clause is formed by affixing ‘அ (*a*)’ to the non-finite verb. In (3), ‘அணில் வந்த\_காரணத்தை அறிய (*anil vandha\_kaaranathai ariya -* to know the reason of squirrel’s arrival)’ is the infinite verbal clause. Sometimes, the infinite form act as adverbial clause.

All the remaining cases such as ‘special verbal participle’ and ‘negative special verbal participle’ comes in between the clause and the verb. (11) is a sample special verbal participle ‘endru’ and it is similar to the English subordinate clause ‘that / about’.

1. இந்த நாட்டில் பெரும்பான்மையோர் **அடிமைத்தன ஒழிப்பை**

*indha naatil perumbaanmaiyor* ***adaimaithana ozhippai***

this in.country majority.people(S) **slavery abolishment (O)**

விரும்புகிறார்.

*Virumbigiraargal*.

wants (V).

Majority people in this country wants the abolishment of slavery.

1. புலிக்கு (*pulikku*, to a tiger / for a tiger)
2. நிலங்களை வாங்கியவர்கள் **பணத்தை இழந்து, நிலத்தையும்**

*nilangalai vaangiyavargal* ***panathai izhandhu nilathaiyum***

lands those.who.bought (S) **money having.lost land**

**இழந்து** தவிக்கிறார்கள்.

***izhandhu*** *thavikiraargal.*

**having.lost (O)** are.suffering (V).

**Having lost the money and land**, those who bought the lands are suffering.

1. அப்போது சம்பந்தர், **தந்தையைக் காணாமல்**

*appodhu sambandhar* ***thandhaiyai kaanaamal***

that.time Sambandhar (S) **father not.having.found (O)**

திகைத்தார்.

*thigaithaar.*

was.shocked (V).

**Not having found the father**, Sambandhar was shocked that time.

1. **பணம் காணாமல் போனதால்** அனைவரும்

***panam kaanaamal ponadhaal*** *anaivarum*

**money missing because.went (O)** everyone (S)

செய்வதறியாது திகைக்கிறார்கள்.

*seivadhariyaadhu thigaikiraargal*.

not.having.known.to.do is.shocked (V).

Everyone is shocked **because the money went missing** and don’t know what to do.

1. சிலர் **இவ்வளவு செழிப்பான இலக்கியம் தமிழில்**

*silar* ***ivvalavu sezhipaana ilakiyam thamizhil***

someone(S) **how rich literature in.tamil**

**இருக்கிறதா என்று** வியந்துபோவார்கள்.

***Irukiradhaa endru*** *viyandhupovaargal*.

**is.there that (O)** will.wonder (V).

Someone will wonder **that how rich literature is there in Tamil**.

1. **Case\_of\_object\_inTamil:**

This column provides the Tamil grammatical terms for object cases. There are totally 5 values in this column. They are ‘2nd vaetrumai 'ஐ', ‘4th vaetrumai 'கு', irandha kaala vinaiyecham, ethir kaala vinaiyecham and idaichol thodar’.

In Tamil, the accusative case is called ‘2nd vaetrumai’. When the ‘Case\_of\_object’ column value contains ‘accusative’, then this column’s value is ‘2nd vaetrumai 'ஐ'’. The dative case is called ‘4th vaetrumai’. When the ‘Case\_of\_object’ column value contains ‘dative, then this column’s value is ‘4th vaetrumai 'கு'’.

The verbal participle clause and negative verbal participle clause are called ‘irandha kaala vinaiyecham’. The infinite verbal clause is called ‘ethir kaala vinaiyecham’.

All the remaining cases such as adverbial phrase, adverbial clause, special verbal participle and negative special verbal participle are called ‘idaichol thodar’ in Tamil.

1. **Animacy\_of\_subject:**

This column defines the animacy of the subject. The values are ‘rational’, ‘animate’ and ‘inanimate’. When the subject is human-being, then the value is ‘rational’. If it is animal, then the value is ‘animate’. If it is a non-living thing, then the value is ‘inanimate’.

In (9), the subject (Sambandhar) is a human-being and hence the value of this column will be ‘rational’.

1. **Animacy\_of\_object:**

This column defines the animacy of the object. The values are ‘rational’, ‘animate’, ‘inanimate’ and ‘none’. When the object is human-being, then the value is ‘rational’. If it is animal, then the value is ‘animate’. If it is a non-living thing, then the value is ‘inanimate’.In case if the object is clausal, then the value should be ‘none’.

In (5), the bold words are clausal object and hence the value of this column will be ‘none’.

1. **Pronoun\_subject:**

This column defines whether the subject is pronoun or not. The values are ‘yes’ and ‘no’. If the subject is a pronoun, then the value is ‘yes’. Else, it will be ‘no’.

In (12), the subject (நான் - I) is a pronoun and hence the value of this column will be ‘yes‘.

1. **நான்** கரையில் நின்று **அவரை** வியக்கின்றேன்.

***naan*** *karaiyil ninru* ***avarai*** *viyakkinren.*

**I** (S) in.shore by.standing **him** (O) wonder (V).

I wonder him by standing on the shore.

1. **Pronoun\_object:**

This column defines whether the object is pronoun or not. The values are ‘yes’ and ‘no’. If the object is a pronoun, then the value is ‘yes’. Else, the value will be ‘no’.

In (12), the object (அவரை - him) is a pronoun and hence the value will be ‘yes’.

1. **Syllable\_of\_subject:**

It contains the no. of words in the subject. The values are finite integers that starts from 1. In (11), the subject is ‘சிலர்’ and that is one word, hence the syllable is 1.

1. **Syllable\_of\_object:**

It contains the no. of words in the object. The values are finite integers and starts from 1. In (11), the object is ‘இவ்வளவு செழிப்பான இலக்கியம் தமிழில் இருக்கிறதா என்று’ and they are 6 words. Hence, the syllable is 6.

1. **Focus:**

This column indicates which part is focused more on a sentence, whether it is subject or object or verb. The values of this column are ‘none’, ‘subject’, ‘subject (emphasized)’, ‘object (emphasized)’, ‘subject (emphasized), object(emphasized)’ and ‘verb (emphasized)’.

In Tamil, particles like ‘தான் (*thaan*)’, ‘கூட(*kooda*)’ and ‘உம் (*um*)’ are affixed to the subject or object inorder to emphasize them.

In Tamil, the part which occurs before verb is considered to be more focused. But, it doesn’t apply for the SOV order, as this is the standard word order of Tamil. Suppose if the order is OSV, it means subject is more focused as it occurs before verb. In (5), the word order is OSV and hence the value of this column is ‘subject’.

If there are no emphasized particles and the order is SOV, then the value is ‘none’. Despite the word order, if the subject or object is emphasized, then the value is either ‘subject (emphasized)’ or ‘object (emphasized)’ accordingly. In case if both the subject and object are emphasized, then the value is ‘subject (emphasized), object(emphasized)’. In rare scenario, verbs are emphasized with particles like 'போ (*po*)', 'விடு (*vidu*)’. In those cases, the value is ‘verb (emphasized)’.

1. **Psych:**

This column indicates whether the verb in a sentence denotes the emotional (psych) meaning or not. This column has only one value ‘psych’. Because all verbs of this dataset denote only the emotional meaning in their sentences and there is no non-emotional or ambiguity occurrences.