Ruhr-Universität Bochum

Linguistic Data Science

Research Project- 2

Word order with psych verbs in Tamil language

Submitted by: Sadhana Muthukumar, Master's student in Linguistic Data science
Supervisors: Simon Masloch, M.A.; Johanna M. Poppek, M.A.
Date: 04.09.2023
Place: Bochum, Germany

Abstract

The word order of a language is based on subject, object and verb. In Tamil, SOV is the standard and the most common word order. This project analyzes the word order pattern for psych verbs in Tamil language. Generally, psych verbs are different from other verbs. These psych verbs can be seen in all languages. 'To be happy', 'to fear', 'to like' are some examples for psych verbs and these verbs represent only emotions and no physical movements. Psych verbs are considered to be peculiar among linguists as they do not follow the rules that has been applied to verbs in the grammar theory. Psych verbs has two main characteristics and they are as follows. 1, In psych verbs, the object can act as an experiencer and the subject act as a stimulus of the emotion. 2, These psych verbs deviate from the preferred word order. The word order deviation of experiencer object (EO) psych verbs has been already proved in German language by (Temme, 2019). Following these facts, this project determines whether the psych verbs of Tamil deviate the word order or not. Due to the dominance of light verb construction in modern Tamil, the availability of EO verbs in the form of main verb is very less. Owing to this limitation, this work has taken both type of psych verbs i.e., experiencer subject (ES) and experiencer object (EO) into account and examined them. Two annotation datasets were created for the analysis. The first one is for psych verbs and the second is for non-psych verbs. The purpose of second dataset is just to compare the word order of psych verbs with non-psych. The analysis was carried out using generalized models through R programming language. It was also analyzed from the native speaker point of view. All these results showed that, the psych verbs in Tamil do not deviate the word order. In addition, SOV word order is more prevalent among psych verbs in Tamil. In fact, it is the common word order for this language.

தொகுப்புரை தமிழில் வருமாறு (Abstract in Tamil)

ஒரு மொழியில் சொல் வரிசை என்பது எழுவாய், செய்படுபொருள், பயனிலை என்ற இந்த மூன்றை அடிப்படையாகக் கொண்டமையும். தமிழ் மொழியில் எழுவாய் முதலிலும், செய்படுபொருள் அதன் பின்னும், இறுதியில் பயனிலையும் கொண்ட வரிசையே மிக பொதுவான வரிசையாகும். தமிழ் மொழியில் உணர்வு வினைச்சொற்கள் கொண்ட சொற்றொடர்களுடைய சொல் வரிசையை பற்றியே இந்த ஆராய்ச்சி பகுப்பாய்வு செய்கிறது. உணர்வு வினைச்சொற்கள் மற்ற வினைச்சொற்களை காட்டிலும் மாறுபட்டு இருக்கிறது. இவ்வகையான வினைச்சொற்கள் எல்லா மொழிகளிலும் காணப்படும். மகிழ், அஞ்சு, விரும்பு போன்றவை உணர்வு வினைச்சொற்களாகும். ஏனெனில் இவை உணர்வை தவிர எவ்வித செயலையும் குறிக்காத.. பொதுவாக மொழியியலாளர்களிடையே இவ்வகை வினைச்சொற்கள் தனித்தன்மை கொண்டதாக பார்க்கப்படுகிறது. காரணம், இவைகள் வினைச்சொற்களின் பொதுவான இலக்கண முறையை பின்பற்றுவுது இல்லை. அதில் முக்கிய தன்மைகள் இரண்டாகும். ஒன்று இவ்வினைச்சொற்களை கொண்ட தொடர்கள் சிலவற்றில் செய்படுபொருள் உணர்வை கொண்டதாகவும், எழுவாய் உணர்வை தரக்கூடியவையாகவும்

[1] My sincere thanks to my supervisors Mr. Masloch and Ms. Poppek for their immense support throughout this project !!

அமையும். இரண்டாவது இவை தொடரின் சொல் வரிசையை மாறுபடுத்தும் தன்மையை கொண்டுள்ளது. இந்த சொல் வரிசை மாற்றத்தை ஜெர்மனி மொழியில் ஏற்கனவே டெம்மி என்பவர் அவரின் ஆராய்ச்சி கட்டுரையில் உறுதி செய்துள்ளார். இதனடிப்படையில் தமிழிலும் இவ்வினைச்சொற்கள் சொல் வரிசையை மாற்றுகிறதா இல்லையா என்பதையே இந்த ஆராய்ச்சி ஆராய்கிறது. இக்காலத்தமிழில் திரிதல் ஆதிக்கம் மிக இருப்பதால் தமிழ் இலக்கண முறையை பின்பற்றும் உணர்வு வினைச்சொற்கள் குறைவாகவே உள்ளது. இதன் காரணத்தினால் எழுவாய், செய்படுபொருள் என்ற இருவகையான உணர்வு வினைச்சொற்களையும் இந்த ஆராய்ச்சி கருத்தில் கொள்கிறது. இவ்வாராய்ச்சிக்காக இரண்டு தொகுப்புகள் உருவாக்கபட்டது. முதல் தொகுப்பு உணர்வு வினைச்சொற்கள் வரும் தொடர்களையும், இரண்டாம் தொகுப்பு செயலை குறிக்கும் வினைச்சொற்கள் வரும் தொடர்களையும் கொண்டதாக அமைந்தது. இரண்டாம் தொகுப்பின் நோக்கம் ஒப்புமைக்காக மட்டுமே. இத்தொகுப்புகளில் உள்ள தொடர்களின் சொல் வரிசையை 'ஜெனரலைஸ்டு மாடல்ஸ்' எனப்படும் மென்பொருள் கருவிகள் மூலம் 'ஆர்' நிரலாக்க மொழி வழியாக பகுப்பாய்வு செய்யப்பட்டது. மேலும், சொல் வரிசைகளை தாய் மொழி என்ற அடிப்படையிலும், தமிழ் மொழியின் இயல்பின் அடிப்படையிலும் ஆய்வு செய்யப்பட்டது. இந்த அனைத்து ஆய்வுகளின் முடிவு என்னவென்றால் தமிழ் மொழியில் உணர்வு வினைச்சொற்களுக்கு சொல் வரிசையை மாற்றும் தன்மை இல்லை. மேலும், இவ்வினைச்சொற்கள் வரும் தொடர்கள் எழுவாய், செய்படுபொருள், பயனிலை என்ற வரிசையையே அதிகமாக கொண்டுள்ளது. இவ்வரிசையே தமிழ் மொழியின் பொதுவான வரிசை என்பதும் குறிப்பிடத்தக்கது.

1. Introduction

Given the psychological properties, verbs can be classified into 2 major groups i.e., psychological (psych) and non-psychological (non-psych) verbs. Former is the one where there is no physical movement involved but only emotions and the later deals with physical movements. (1) and (2) are the examples for psych and non-psych verbs respectively. In (1), the verb 'fears' denotes a certain emotional state of Clara, whereas in (2) the verb 'moved' denotes the physical motion.

(1) Clara *fears* the noise.

(2) He *moved* the chair.

In psych verb, one of the arguments i.e., subject or object is in certain emotional state. The argument being in the emotional state is called experiencer and the other which stimulates the emotional state is called stimulus. This further leads to 2 sub types namely Experiencer Subject (ES) and Experiencer Object (EO). (3) and (4) are the examples for ES and EO respectively. In (3), the argument which experiences the emotion is a subject 'I'. In (4), the argument which experiences the emotion is an object 'everyone'.

(3) *I* love nature.

(4) His jokes amused *everyone*.

Since decades, the psych verbs particularly EO verbs are quite peculiar among linguistic researchers due to their special nature. According to (Belletti & Rizzi, 1988), (Landau, 2010)

and many others, the psych verbs behave syntactically and semantically unusual, in contrast to non-psych verbs. The work of (Temme, 2019) proved that the dative EO psych verbs in German language prefers first position in the sentence. In fact, the analysis of syntactic pattern of German EO verbs by (Masloch et al., 2021) is being the main motivation for the existence of this work. Since there is no relevant research work on psych verbs of Tamil so far, that again encouraged to initiate this work. This work considered both ES and EO psych verbs for the word order analysis.

2. Background

2.1. Tamil and its grammar

Tamil (Thamizh) is one of the most oldest languages in the world, which is surviving till date. Tamil is dated to approximately 200 BCE (Steever, 2019). Yet, the origin of this language is still unknown. It belongs to the family of Dravidian languages and is natively spoken by the people of Tamil Nadu and Puducherry in India and in some northern parts of Sri Lanka. Being the oldest language, it can be classified into three stages such as Old Tamil (c .300 BCE to 700 CE) , Middle Tamil (700 to 1600 CE) and Modern Tamil (1600 CE to the present) (Steever, 2019). This work deals only with the written form of formal modern Tamil.

Tholkaapiyam is the most ancient literature of Tamil dated to 200 BCE and it is the source of Tamil grammar. Nannool dated to 1000 CE derived from tholkappiyam is considered as the second biggest source of Tamil grammar. In Tamil, there are 12 vowels[2] and 18 consonants[3]. They form the basic letters.

The vowels and consonants combined each other and form the compound letters. There are 216 compound letters i.e., 12 vowels * 18 consonants = 216 compound letters. Also, there is 1 special letter ஃ (*ak*). Therefore, there are totally 247 letters in Tamil. 12 vowels + 18 consonants + 216 compound letters + 1 special letters = 247 letters.

2.2. Rules of Tamil grammar

Tamil is agglutinative in nature and explicitly suffix the particles at its end. Though there are plenty of rules in Tamil grammar, only certain important rules that is necessary for this work are explained here (Arden, 1891), (இராச திருமாவளவன், n.d.).

---

[2] Vowels

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
a	*aa*	*e*	*ee*	*u*	*oo*	*ye*	*yae*	*i*	*o*	*oa*	*ow*

[3] Consonants

க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்
ik	*ing*	*ich*	*in*	*it*	*in*	*ith*	*inth*	*ip*	*im*	*iy*

ர்	ல்	வ்	ழ்	ள்	ற்	ன்
ir	*il*	*iv*	*izh*	*il*	*ir*	*in*

Rule 1: The verb always stands at the end and concludes the sentence. In (5), the verb *படித்தான்* (*padithaan* - learnt) lies at the end.

(5) *பையன் பாடத்தை படித்தான்.*

paiyan paadathai padithaan.
boy lesson learnt.
Boy learnt the lesson.

Rule 2: The verb should agree with the subject. Tamil verbs indicate the tense, the gender of subject including its honor (formal or informal) and the number (singular or plural) of subject. When the subject is a first person pronoun *நான்* (*naan* - I) or second person *நீ* (*nee* - you), it is not necessary to include the subject in sentence as the verb itself defines them. In (6), the verb *வியந்தேன்* (*viyandhen – I wondered*) indicates that the subject is a first person *நான்* (*naan* - I) and hence it is not necessary to mention the subject.

(6) *வியந்தேன்* = *விய* + *ந்து* + *ஏன்*
= *viya* + *indhu* + *yaen*
(verb lemma) (past tense) (first person singular)
= *viyandhen* - I wondered

Rule 3: The nouns are classified into 2 major classes such as rational and irrational. Rational nouns denote human beings and Irrational nouns denote animals and inanimate things.

Rule 4: There are no prepositions in Tamil, but only postpositions. These postpositions affix with nouns and pronouns at the end, inflect them in order to give a new meaning. In (7), the postposition *இல்* (*il*) suffixed with the noun and gives a new meaning.

(7) *பசுவில்* = *பசு* + *இல்*
= *pasu* + *il*
(noun) (postposition)
= cow + in
pasuvil = in a cow

Rule 5: There are no articles in Tamil.

Rule 6: The subordinate clause which expresses purpose, reason and condition always precedes the main clause. In (8), the reason (because) occurs at the initial position and the main clause comes after it.

(8) *இதனால் 1_லட்சத்து_80_ஆயிரம் பயணிகள் தங்கள்*
idhanaal 1_lachathu_80_aayiram payanigal thangal
because.of.this1.lakh.80.thousand passengers their

இடங்களுக்குச் செல்ல முடியாமல் தவித்தனர்.
idangalukku sella mudiyaamal thavithanar.
places to.go not.able.to were.suffered.

Because of this, 1 lakh 80 thousand passengers were suffered by not able to go to their places.

Rule 7: There should be only one finite verb in a sentence and the other verbs are in the form of non-finite. These non-finite verbs usually end the subordinate clause. In (9) there is only one finite verb *வியந்தனர்* (*viyandhanar* - wondered) and the other verb *பார்த்து* (*paarthu* – having seen) is in non-finite verbal participle form. This non-finite verb ends the subordinate clause.

(9)	ஒரு	சிறுவன்	அதில்	கலந்து _கொண்டதைப்	பார்த்து
	oru	*siruvan*	*adhil*	*kalandhu_kondadai*	*paarthu*
	one	little.boy	in.it	having.participated	having.seen

எல்லோரும்	வியந்தனர்.
ellorum	*viyandhanar.*
everyone	wondered.

Everyone wondered on seeing one little boy having participated in it.

Rule 8: There are 8 cases in Tamil and these cases affixed with nouns or pronouns.

Case1: Nominative. In this case, it is the noun itself. For e.g., *நரி* (*nari* - a Jackal)

Case2: Vocative - *ஏ* (*yae*)

(10)	நரி	+	ஏ	=	நரியே
	nari	+	*yae*	=	*nariyae*
	jackal	+	O	=	O Jackal

Case3: Genitive - *இன்* (*in*)

(11)	புலி	+	இன்	=	புலியின்
	puli	+	*in*	=	*puliyin*
	tiger	+	of	=	of a tiger

Case4: Dative – *கு* (*ku*)

(12)	புலி	+	கு	=	புலிக்கு
	puli	+	*ku*	=	*pulikku*
	tiger	+	to	=	to a tiger

Case5: Accusative - *ஐ* (i*)*

(13)	புலி	+	ஐ	=	புலியை
	puli	+	*i*	=	*puliyai* - a tiger

Case6: Locative – *இல்* (*il*)

(14) புலி + இல் = புலியில்

puli + il = puliyil
tiger + in = in a tiger

Case7: Instrumental - ஆல் (*aal*)

(15) புலி + ஆல் = புலியால்
puli + *aal* = *puliyaal*
tiger + by = by a tiger

Case8: கண் (*kan*). It is specific only to Tamil and there is no English equivalent for this case.

3. Study of word order through annotation

Tamil is a free word order language where both SOV and OSV is allowed with one strict rule that verb should be at the end always. Poetic literatures of Tamil are an exception, where verb can come in any of the three positions so as to enhance the aesthetic. All the other forms except poetics should have final verb position. There are several factors which leads to the scrambling of word order including case of object, focus , length of subject and object etc. In order to determine the word order and its factors, it is necessary to annotate the sentences. As there was no annotated dataset available for Tamil psych verbs, creating manual annotation following the rules of Tamil grammar became one of the major parts of this work. As a first step, 10 psych verbs of Tamil were collected. 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), வெறு (*veru* - to hate). For comparison purpose, two non-psych verbs such as படி (padi - to read / to study), தேடு (thaedu - to search / to find) were also taken.

3.1. Dataset

Leipzig corpora (*Download Corpora Tamil*, n.d.) is the data source for this work, from where the sentences containing these verbs were retrieved. This corpus is a set of modern written Tamil collected from community, news, newscrawl, web and Wikipedia for the years of 2011, 2016, 2017, 2019, 2020 and 2021. This work also focused on the formal written Tamil[4]. This corpus includes both Indian and Sri Lankan Tamil[5]. Sentences were extracted from this corpus with the help of Pandas tool through Python programming language. The initial plan was to annotate 200 sentences for each verb. Due to data constraint, the no. of sentences varies for every verb. Especially, the count for களி (*kali* - to ecstasize) and நெகிழ் (*negizh* - to delight) is very less.

[4] In modern Tamil, the way words are uttered in spoken is different from the way it is written. For e.g., the word அவர்கள் (*avaragal*) is used in written to express 'they', but the same has been uttered as அவங்க (*avanga*) in spoken. Apart from the utterance variation, all the other grammatic, syntax and semantic concepts are same between them.

[5] Tamil has various geographical dialects, yet the major distinction can be found between the dialects of Indian Tamil and Sri Lankan Tamil. However, the grammatical concepts are exactly similar between them other than dialect and usage of certain different words.

3.2. Rules for annotation

Certain limitations were followed to perform the annotation. Sentences where there is no subject or object were excluded. The usage of light verbs to express the psych properties is quite high in modern Tamil. In light verb construction, the actual verb denoting the psych meaning comes in a noun form and is followed by the light verb. Yet, both combines and deliver the same psych meaning.

These light verbs cannot standalone and give a complete meaning. Among these 10 verbs, only விரும்பு (*virumbu* - to like) has no light verb construction in any of its sentences.

The common light verbs used to express the psych properties are அடை (*adai* – to attain), அளி (*ali* – to give), திளை (*thilai* – to indulge), ஆழ் (*aazh*, to immerse) and செய் (*sei* – to make). In (16), the actual psych verb occurs in noun form வியப்பை (*viyappai,* wonder) and the light verb அளித்தன (*alithana* – gave) comes after it.

(16)	அந்தப்	பெண்ணின்	பிடிவாதமும்,	வீம்பும்	எனக்கு	வியப்பை
	andha	*pennin*	*pidivaadhamum,*	*veembum*	*enakku*	*viyappai*
	that	women's	adamant,	stubbornness	me	wonder

அளித்தன.
alithana.
gave.
That women's adamant and stubbornness gave me wonder.

The light verb usage is very vast in EO type than ES. (16) is also an EO with light verb where the experiencer is an object எனக்கு (*enaku*, me). Certain verbs like மகிழ் (*magizh* - to be happy), நெகிழ் (*negizh* - to delight), அஞ்சு (*anju* -to fear), வருந்து (*varundhu* - to be sad) are having their own EO form. The particle called இ (*e*) will be suffixed to the verb lemma in order to denote the EO form. With these EO form of verbs, light verb construction is absent. But still, light verb construction is possible among these verbs without their EO forms. Yet, the count of light verb construction is high for these verbs though they have EO forms.

The usage of dative subjects is also very common in modern Tamil (Murugaiyan, 2011). According to Tamil grammar, the subject should be nominative without any particle inflection. But today, subject in dative form is widely used and this violates the subject verb agreement. When the subject is dative, the verb comes in the form of noun and still delivers the meaning of verb. In (17), the subject எனக்கு (*enakku* - I) is in dative and the verb is in noun form வருத்தம் (*varutham* - sad). Though, they deliver the verb meaning of i.e. to be sad.

(17)	எனக்கு	அவள்மீது	வருத்தம்.
	enakku	*avalmeedhu*	*varutham.*
	I	her.on	sad (noun).

I feel sad about her.

Through this work, it was identified that the light verbs and dative subjects are used only for the psych verbs of modern Tamil. They are not used in any non-psych verbs. Sentences containing light verbs and dative subjects were excluded from annotation work, as they

modify the original form of psych verbs which actually plays a central role in this work. Sentences with passive voice were also excluded.

Sentences containing both subject, object as well as the psych verb as main finite verb are selected for the annotation. Annotation work determined the type of psych verb, word order and factors affecting word order such as case of subject, case of object, animacy of subject, animacy of object, pronoun of subject, pronoun of object, subject syllable, object syllable and focus. The final annotation file has totally 1412 annotated sentences in the form of rows and factors in the form of columns. In this annotation, subjects with and without phrase are considered as subject. The finite verbs are considered as verb. The object cases are broadly classified into clausal and non-clausal. The clauses that end with non-finite verb are considered as clausal object. The phrases without non-finite verb denoting the object are considered as non-clausal object.

(18)	*பல்லாயிரக்கணக்கானவர்கள்*	*இந்தப்*	*புடவையைப்*	*பார்த்து*	*வியக்கின்றனர்.*
	pallaayirakanakaanavargal	*indha*	*pudavaiyai*	*paarthu*	*viyakkinranar.*
	thousands.of.people	this	saree	having.seen	are.wondering.

Thousands of people are wondering by having seen this saree.

Psych verbs are generally defined in terms of experiencer and stimulus, where the experiencer can be a subject and the stimulus can be an object or vice versa. In (18), the subject is *பல்லாயிரக்கணக்கானவர்கள்* (*pallaayirakanakaanavargal* – thousands of people) and verb is *வியக்கின்றனர்* (*viyakkinranar* – are wondering). Here, the subject is experiencing the emotion and hence subject is an experiencer. The clause *இந்தப் புடவையைப் பார்த்து* (*indha pudavaiyai paarthu* - having seen this saree) is acting as a stimulus. In the sense, this clause stimulates the experiencer in order to wonder i.e., thousands of people are wondering by having seen this saree. Therefore, the entire clause has been considered as an object. In this annotated data, several sentences containing clausal object and few are having non-clausal object. Among these clausal objects, verbal participle case is more prevalent and is followed by special verbal participle. In (18), the word *பார்த்து* (*paarthu*, having seen) is the verbal participle form of verb *பார்* (*paar* - to see). Verbal participles namely *கண்டு* (*kandu,* having seen), *கேட்டு* (*kaetu* – having heard) and special verbal participle especially *என்று* (*endru*), *என* (*ena*) are very common and more frequently used with Tamil psych verbs.

Apart from subject and object, the other additional factors are animacies and pronouns of subject and object. The frequent animacy of subject and object is rational and none respectively. The occurrence of pronouns is very less in both subject and object.
The other main factors are focus and syllables. Generally, in Tamil the most common way to focus elements is by suffixing the particles like *தான்* (*thaan*), *ஏ* (*yae*) or *ஓ* (*o*) to them. The least way is through changing the word order. In SOV, either nothing is focused or the elements are focused using particles. In OSV, either the subject is focused without particles or object is focused with particles. It is to be noted that the subject is not getting focused everytime in OSV and it purely depends on the sentence sematics.
The total no. of words is considered as syllable. The syllables were counted for subject and object.

4. Analyzing the word order of psych verbs using Generalized models

The analyzes were implemented using generalized linear and generalized mixed models through R programming in order to make it more effective and efficient. Before proceeding with the models, some initial pre-processing of dataset was performed such as removing the unwanted columns, converting categorical variables into factors and numerical variables into logs for getting the good result.

Model 1: Generalized Linear Model (GLM)

When the response variable is a categorical variable, then GLM will be used. In GLM, the categorical response variable is modeled as a function of one or more predictors that can be either categorical or continuous variables. GLM observes the probability of a single event of response variable.

Initially two GLM models were created. One is a complex and the other is a simple model. Phrasal_subject and Clausal_object are not added as predictors in both the models as they are correlating with the Case_of_subject and Case_of_object respectively. The complex model has all the predictors. The simple model has predictors other than animacy and pronouns, since adding these variables decreasing the significance of the model as well as they doesn't have strong effect on the word order. These models were compared using Anova test. The test result talks about the efficiency of complex model. If p-value is <0.05, then the complex model is significant or else it is not significant. The p-value for this test is 0.23. This shows that the complex model is not significant. Hence, the simple model called 'tamil.psych.glm1' has been chosen.

‘tamil.psych.verb.glm1 <- glm(Word_order ~ EO_ES + Case_of_subject + Case_of_object + log_syllable_subject + log_syllable_object + Focus, family ="binomial", data = tamil.psych.verb.df)’

The response variable is a Word_order that is binomial as it has 2 values namely 'SOV' and 'OSV'. The predictors are EO_ES, Case_of_subject, Case_of_object, log_syllable_subject and log_syllable_object and Focus.

Table 1. Summary of model 1

```
Call:
glm(formula = Word_order ~ EO_ES + Case_of_subject + Case_of_object +
    log_syllable_subject + log_syllable_object + Focus, family = "binomial",
    data = tamil.psych.verb.df)

Coefficients:
                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                      5.97223    0.54002  11.059  < 2e-16 ***
EO_ESEO                                         -0.03413    0.93518  -0.036 0.970890
Case_of_subjectnominative (phrasal)              0.28168    0.45035   0.625 0.531669
Case_of_objectspecial verbal participle         -2.34947    0.32444  -7.242 4.43e-13 ***
Case_of_objectaccusative                        -1.37038    0.46864  -2.924 0.003454 **
Case_of_objectadverbial clause                  -1.59350    0.69254  -2.301 0.021395 *
Case_of_objectadverbial phrase                  -0.56394    1.16627  -0.484 0.628712
Case_of_objectdative                            -2.36454    0.98054  -2.411 0.015888 *
Case_of_objectinfinite verbal participle        -0.28451    0.45625  -0.624 0.532908
Case_of_objectparticiple noun clause            -1.04138    0.60311  -1.727 0.084224 .
log_syllable_subject                             1.18945    0.35189   3.380 0.000724 ***
log_syllable_object                             -1.95472    0.22261  -8.781  < 2e-16 ***
Focusobject (emphasized)                        -1.73825    0.31058  -5.597 2.18e-08 ***
Focussubject                                    -8.33712    0.69283 -12.033  < 2e-16 ***
Focussubject (emphasized)                       -2.06983    0.37064  -5.585 2.34e-08 ***
Focussubject (emphasized), object (emphasized)  -1.83881    1.21504  -1.513 0.130184
Focusverb (emphasized)                          13.81377  616.13613   0.022 0.982113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1888.83  on 1411  degrees of freedom
Residual deviance:  476.69  on 1395  degrees of freedom
AIC: 510.69

Number of Fisher Scoring iterations: 14
```

Table 1. shows the result of model 1. The coefficients provide the intercept and slopes values. Intercept is the Word_order and slopes are the predictors. These values are the log odds (logit) that has to be converted into probability. The category shown to the right is taken as the reference level for the response variable. Thus, SOV is the reference level for Word_order. The category which occurs first in the alphabetical order is taken as the reference level for predictors. The reference level for EO_ES and case of object has been changed to ES and verbal participle clause as they are very abundant.

The positive slope is the indication of increase in the SOV order and the negative slope decreases the SOV order. For e.g., subject with nominative (phrasal) category has positive slope. It means the odds of observing SOV increases when the subject is phrasal. The object with special verbal participle has a negative slope meaning the odds of observing SOV order decreases for it.

Despite of positive and negative slopes, it is worth to consider only the predictors those having significant p-values.When the p-value <0.05, then the elements are significant and has a strong affect with the word order[6]. Predictors such as special verbal participle, log_syllable_subject, log_syllable_object, focus with object(emphasized), subject, subject (emphasized) are highly significant. It means they are having a very strong effect towards word order. The intercept is also highly significant in this model.

[6] *** means highly significant, ** is medium and * is less significant.

The probability of SOV is 100% and OSV is 0% when verb is ES, subject is nominative, object is verbal participle clause and nothing is focused. This shows that the ES and verbal participle clause prefers only SOV when length and focus is not considered. If focus comes into picture, the SOV decreases significantly when subject is focused as well as when object is focused with particles. The subject with high syllable length prefers SOV order and the object with high syllable length prefers OSV. Special verbal participle also has a great tendency for OSV. In addition, accusative, dative and adverbial clause are also having OSV tendency. But their impact is minimum. Overall, the probability of observing SOV is higher than OSV in this model.

Evaluation of model 1

The concordance index, also known as C-index is a metric used to evaluate the logistic regression model. It takes all the sentences whose word order has been correctly predicted by the model. Within them, it makes all the possible pairs of sentences containing SOV and OSV order. The final C value will be the ratio of the no. of pairs where predicted probability of SOV is higher than OSV to the ratio of total no. of pairs.

When the C value is >= 0.9, model has outstanding discrimination. If $0.8 \leq C < 0.9$, then excellent discrimination. If $0.7 \leq C < 0.8$, acceptable discrimination. if $C = 0.5$, then no discrimination. As the C value for this model is 1, it shows that the model discriminates well and is a good fit for the data.

Model 2: Generalized Linear Mixed Model (GLMM)

Generally, there are chances that two or more data points are connected to each other. For e.g., in a survey data, two data points might come from the same person and these data points work in a similar way as they belong to the same person. Independent assumptions are violated if data points are dependent. For Linear Model (LM) and Generalized Linear Model (GLM), data points are assumed to be independent of each other. In model 1 (GLM), the independent assumption has been violated as the data points are dependent to each other in terms of verb. Even though the verb was not added as a predictor in model1, this assumption is still not satisfied.

In fact, the sentences were collected based on the verbs for this work. Multiple sentences belong to the same verb. Those sentences that contains the same verb might have a similar grammatical syntax and functionality. It might be plausible that verb has an influence on the word order. Hence, it is necessary to include the verb in a model in order to determine whether it affects the word order or not. For e.g., SOV is very high for the verb தவி (*thavi* - to suffer) and மகிழ் (*magizh* - to be happy). Besides, it is important to determine whether the effect of special verbal participle, syllables and focus are still hold even after the word order are conditioned on verb. It is impossible to apply LM and GLM model when verb is included. As the verb act as a dependent cluster, it will violate the independent assumption. At this point GLMM comes into picture. It is used for analyzing the data that are dependent to each other. In GLMM, the predictors to which data points are dependent are known as 'random effects'. The other predictors are called 'fixed effects'. The random effect should be categorical, while they account for the cluster dependency and the fixed effect can be categorical or numerical.

‘tamil.psych.verb.glmm <- glmer(Word_order ~ EO_ES + Case_of_subject + Case_of_object + log_syllable_subject + log_syllable_object + Focus + (1 | Verb), family ="binomial", data = tamil.psych.verb.df, nAGQ =0)’

This model contains 1 random effect component i.e., (1 | Verb) and fixed effects as Case_of_subject + Case_of_object + log_syllable_subject + log_syllable_object + Focus. The symbol '|' means 'conditioned on'. The expression (1 | Verb) tells that the word order is conditioned by the verb as random intercept and in this model. In a way, this model is a random intercept model. The random slope estimates the variation between verbs and this variation is not relevant for this analysis. So, there is no random slope for this model.

Table 2. Summary of model 2

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
  nAGQ = 0) [glmerMod]
 Family: binomial  ( logit )
Formula: Word_order ~ EO_ES + Case_of_subject + Case_of_object + log_syllable_subject +
    log_syllable_object + Focus + (1 | Verb)
   Data: tamil.psych.verb.df

     AIC      BIC   logLik deviance df.resid
   485.4    579.9   -224.7    449.4     1394

Scaled residuals:
    Min      1Q  Median      3Q     Max
-9.4644 -0.0374  0.0425  0.1436  7.8952

Random effects:
 Groups Name        Variance Std.Dev.
 Verb   (Intercept) 0.7504   0.8663
Number of obs: 1412, groups:  Verb, 10

Fixed effects:
                                          Estimate Std. Error z value             Pr(>|z|)
(Intercept)                                6.2379     0.6494   9.605 < 0.0000000000000002
EO_ESEO                                   -0.3046     1.1470  -0.266             0.790591
Case_of_subjectnominative (phrasal)        0.1605     0.4751   0.338             0.735570
Case_of_objectaccusative                  -1.4352     0.6819  -2.105             0.035319
Case_of_objectadverbial clause            -1.8890     0.7531  -2.508             0.012135
Case_of_objectadverbial phrase            -1.0741     1.2094  -0.888             0.374506
Case_of_objectdative                      -3.0029     1.0061  -2.985             0.002840
Case_of_objectinfinite verbal participle  -0.6814     0.5829  -1.169             0.242394
Case_of_objectparticiple noun clause      -1.4134     0.6898  -2.049             0.040460
Case_of_objectspecial verbal participle   -2.4090     0.3846  -6.264        0.000000000376
log_syllable_subject                       1.3744     0.3714   3.701             0.000215
log_syllable_object                       -1.9353     0.2337  -8.282 < 0.0000000000000002
Focusobject (emphasized)                  -1.9566     0.3269  -5.986        0.000000002150
Focussubject                              -8.6528     0.7175 -12.060 < 0.0000000000000002
Focussubject (emphasized)                 -2.0910     0.3989  -5.241        0.000000159535
Focussubject (emphasized), object (emphasized)  -1.9506     1.2503  -1.560             0.118731
Focusverb (emphasized)                    16.0745   935.8804   0.017             0.986296

(Intercept)                               ***
EO_ESEO
Case_of_subjectnominative (phrasal)
Case_of_objectaccusative                  *
Case_of_objectadverbial clause            *
Case_of_objectadverbial phrase
Case_of_objectdative                      **
Case_of_objectinfinite verbal participle
Case_of_objectparticiple noun clause      *
Case_of_objectspecial verbal participle   ***
log_syllable_subject                      ***
log_syllable_object                       ***
Focusobject (emphasized)                  ***
Focussubject                              ***
Focussubject (emphasized)                 ***
Focussubject (emphasized), object (emphasized)
Focusverb (emphasized)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let's look at the output of random effect from Table 2. It has a term called Std.Dev that is nothing but the standard deviation. Generally, GLMM estimates the variation of random effects around the intercept. This variation is provided in terms of standard deviation.

The standard deviation of the verb is 0.87 and it is extraordinarily less compared to the value of intercept which is 6.24. When the standard deviation is lesser than the intercept, then the random effect doesn't have great influence with the response variable. Since it is very less, this indicates that the verb has literally no influence with the word order. The intercept is highly significant. Whatever predictors are highly significant in model 1, the same are highly significant in this model too. It means these predictors are still holding their effect even after adding the verb. Again, the probability of SOV is 100% and OSV is 0% with the reference levels. The results of model 1 and model 2 are almost similar owing to the null effect of verb.

Evaluation of model 2

The C value with random effect is 1 and it is same as model 1. It means this model also discriminates well. This again proves that the addition of random effect has not created any impact.

5. Analysis of word order for non-psych verbs

Model 2 has highlighted that the psych verbs are not having any influence with the word order. At this point, it is important to know whether the non-psych verbs are having any influence with the order. So, one GLMM model has been created for non-psych verbs. But only 2 non-psych verbs were considered since it is required only for the comparison purpose.

‘tamil.non.psych.verb.glmm <- glmer(Word_order ~ Case_of_subject + Case_of_object + log_syllable_subject + log_syllable_object + Focus + (1 | Verb), family ="binomial", data = tamil.non.psych.verb.df, nAGQ = 0)’

It consists of 1 random effect component and fixed effects such as (1 | Verb) and Case_of_subject + Case_of_object + log_syllable_subject + log_syllable_object respectively. There is no ES and EO for non-psych and so they were excluded. Also, there will be no clausal object for non-psych and cases of object are accusative, locative and special verbal participle only.

Table 3. Summary of non-psych verb model

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
  nAGQ = 0) [glmerMod]
 Family: binomial  ( logit )
Formula: Word_order ~ Case_of_subject + Case_of_object + log_syllable_subject +
    log_syllable_object + Focus + (1 | Verb)
   Data: tamil.non.psych.verb.df

     AIC      BIC   logLik deviance df.resid
   201.0    244.8    -89.5    179.0      388

Scaled residuals:
     Min       1Q   Median       3Q      Max
-18.1668  -0.0337   0.1138   0.2713  16.5984

Random effects:
 Groups Name        Variance Std.Dev.
 Verb   (Intercept) 0.21     0.4582
Number of obs: 399, groups:  Verb, 2

Fixed effects:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     5.4100     0.8247   6.560 5.38e-11 ***
Case_of_subjectnominative (phrasal)            -1.1404     0.7992  -1.427   0.1536
Case_of_objectlocative                          0.3240     0.6617   0.490   0.6244
Case_of_objectspecial verbal participle         0.9019     0.7873   1.145   0.2520
log_syllable_subject                            0.7087     0.5565   1.273   0.2029
log_syllable_object                            -2.0964     0.3651  -5.742 9.38e-09 ***
Focusobject (emphasized)                       -2.3769     0.5856  -4.059 4.93e-05 ***
Focussubject                                   -8.0438     1.1814  -6.809 9.84e-12 ***
Focussubject (emphasized)                      -1.3725     0.7109  -1.931   0.0535 .
Focussubject (emphasized), object (emphasized)  14.2805  1278.7390   0.011   0.9911
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cs__() Cs_f_b Cs__vp lg_syllbl_s lg_syllbl_b Fcsb(mphszd) Fcssbj Fcss(mphszd)
Cs_f_(phrs) -0.214
Cs_f_bjctlc -0.058 -0.045
Cs_f_bjctvp  0.086 -0.134  0.106
lg_syllbl_s -0.024 -0.834 -0.008  0.082
lg_syllbl_b -0.852  0.189  0.018 -0.183 -0.021
Fcsb(mphszd) -0.382 -0.023 -0.291 -0.041  0.114       0.321
Focussubjct -0.442  0.130 -0.036 -0.061 -0.024       0.440       0.226
Fcss(mphszd) -0.306  0.060  0.036 -0.016 -0.026       0.249       0.213       0.167
F(mphs),o()  0.000  0.000  0.000  0.000  0.000       0.000       0.000       0.000  0.000
```

Let's look at the output of random effect. The standard deviation of the verb is 0.46 that is far less compared to the intercept which is 5.41. Since the variation of verb is far lesser than the intercept, the non-psych verb also has no influence towards the word order (not even minimum influence). This might be owing to the fact that there are only 2 non-psych verbs. Perhaps if the verbs are more, they might be a chance for variability.

The intercept is highly significant. Only 2 predictors namely syllable of object and focus are significant. Unlike psych verbs, the syllable of subject has no effect with non-psych verbs and it perhaps due to less data for non-psych verbs. Here also, the lengthy object and focused subject prefers OSV.

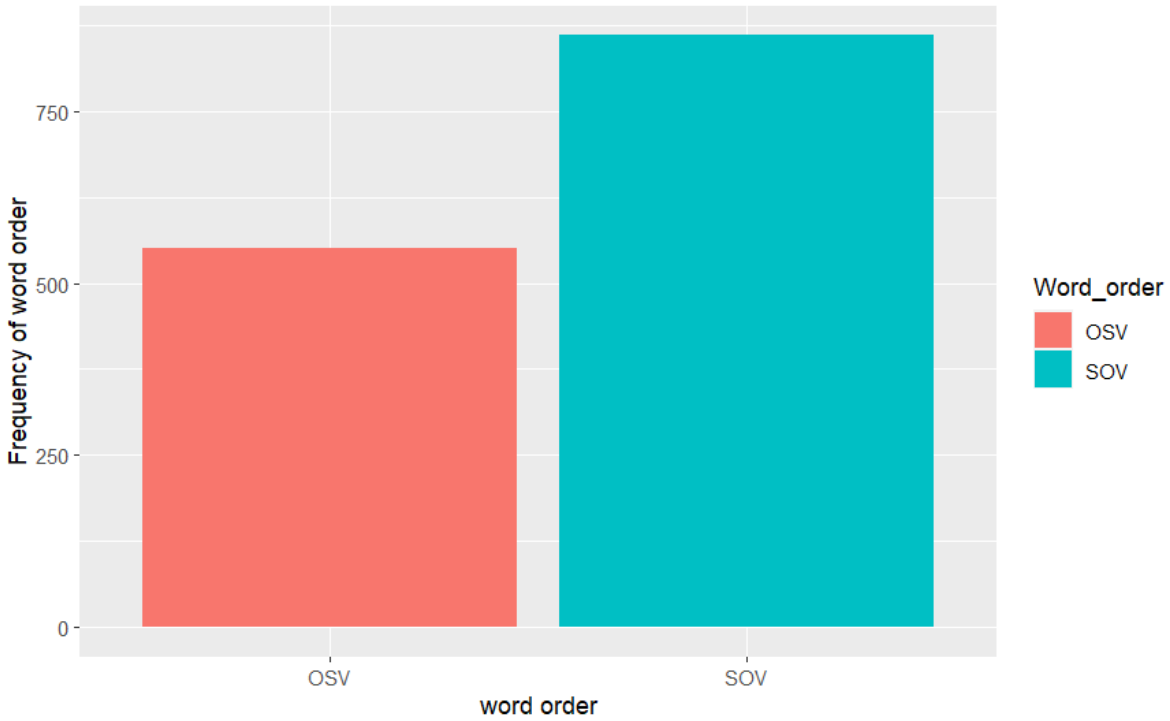
The probability of SOV is 100% and OSV is 0% with the reference levels nominative subject and accusative object. But after considering the syllables and focus, the word order tends to vary between SOV and OSV.

6. Analysis from native speaker perspective

In both model 1 and model 2, predictors namely special verbal participle, syllable of subject, syllable of object and focus holds a strong effect with the word order. Among these, syllable of subject alone prefers SOV when it increases. And the remaining prefers OSV. Yet, the overall probability is far high for SOV and very less i.e., almost 0% for OSV in both the models. It is because, the reference levels are verbal participle clause and 'none' as focus and all these greatly prefers SOV. So, the reference levels has been changed to special verbal participle while it prefers OSV and tried. The overall probability for SOV is 97 % and OSV is 3 % and SOV is still high. Finally, the reference level is changed to 'subject' as focus. This time, the overall probability for SOV is 9 % and for OSV, it is 91 %. This showed that the focus with 'subject' has a very strong effect for OSV. It is to be noted that the syllable of both subject and object were continuously significant in all these times. Hence according to these models, 3 things i.e., lengthy subject, lengthy object and focus with subject determines the word order. When the subject is focused especially without particles, it prefers OSV. When the object is lengthy, it also prefers OSV. But when the subject is lengthy, it prefers SOV order.

One cannot make a decision based on the analysis of models alone. Therefore, the word order was further analyzed from the native speaker point of view. In the annotated dataset, 61% of sentences containing SOV word order and 39% containing OSV. Fig 1. shows the overall count of SOV vs OSV.

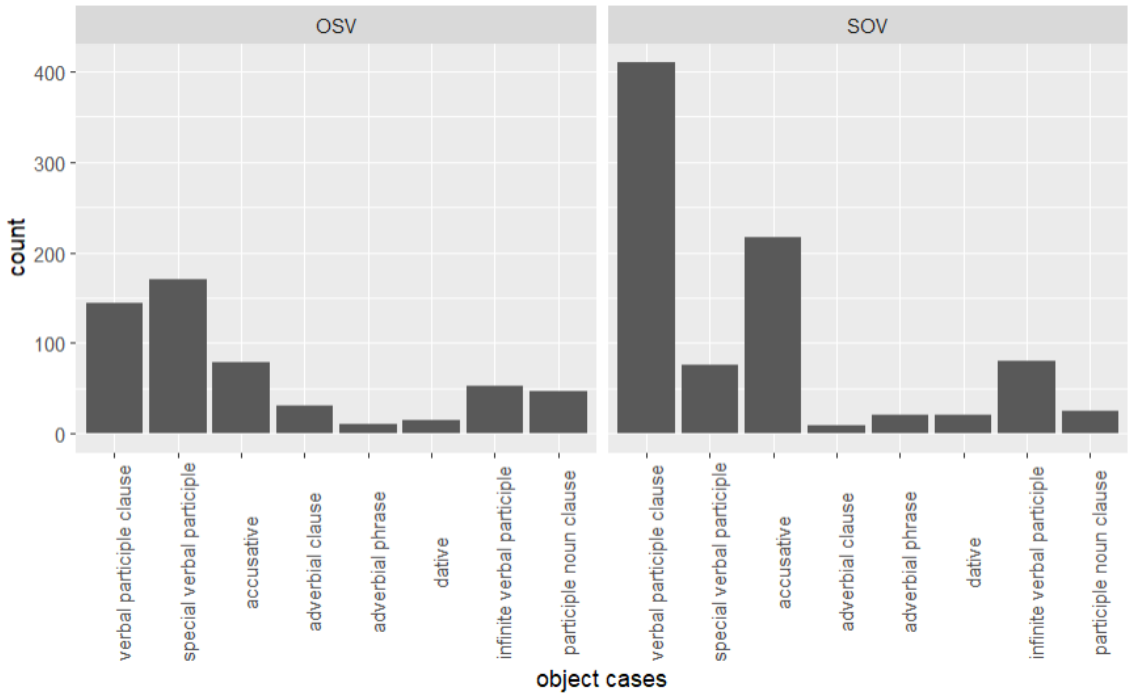
Fig 1. Frequency of SOV vs OSV



Then, all the factors were analyzed one after the other. Among these factors, the case of object, syllable of subject, syllable of object and focus were considered. Remaining factors

especially EO/ES, case of subject, animacy and pronouns were excluded as they don't have any impact. Fig 2. Shows the frequency of cases of object in each SOV and OSV.

Fig 2. Frequency of cases of object



The verbal participle clause cannot be taken for evaluation as it is abundant in both SOV and OSV. The special verbal participle is quite high in OSV and low in SOV. All the other factors are not having distinct variation across the word orders. For analysis purpose, one sentence containing special verbal participle with OSV order has been taken and it is (19).

19 அவள் எந்த பாக்கியசாலிக்கு மருமகளாக வாய்க்கப்_ போகிறாளோ
aval endha bagyasaalikku marumagalaaga vaaikka_pogiraalo
she which lucky.person daughter.in.law going.to.become

என்று சாமி வியந்தான்.
endru saami viyandhaan.
that saami wondered.
Saami wondered that to which lucky person she is going to become daughter in law.

Even when the subject 'Saami' occurs first and followed by the special verbal participle clause, the sentence still gives the same meaning. (20) is the same sentence but with SOV word order.

20 சாமி அவள் எந்த பாக்கியசாலிக்கு மருமகளாக
saami aval endha bagyasaalikku marumagalaaga
saami she which lucky.person daughter.in.law

வாய்க்கப்_போகிறாளோ என்று வியந்தான்.
vaaikka_pogiraalo endru viyandhaan.
going.to.become that wondered.
Saami wondered that to which lucky person she is going to become daughter in law.

Generally in Tamil, there is no preference for the word order based on case of object, though the occurrence of special verbal participle is 31 % in OSV and just 9 % in SOV.

Coming to syllables, the lengthy object prefers OSV. Table 4. Shows that the average length of object is far higher than subject in OSV. The main reason for lengthy object to prefer OSV is to avoid the sentence complexity and to give a clear meaning. This is not compulsory, while the lengthy object still occurs in SOV. But it is to be noted that, the lengthy subject is more prevalent in SOV. Despite of object length, the lengthy subject should occur first. This is further proved in our annotated data.
There is no lengthy subject i.e., > 7 in OSV and it can be seen in only in SOV.

Table 4. Word order vs Syllable of subject and object

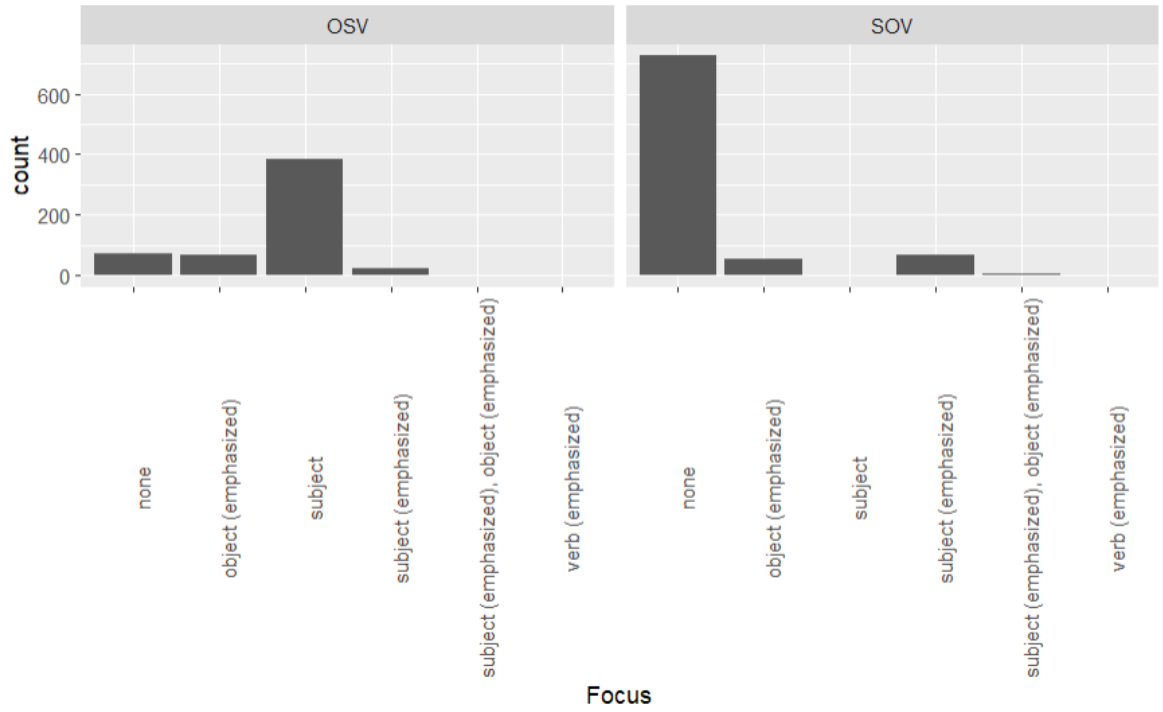
A tibble: 2 × 3

Word_order <fctr>	Syllable_of_subject <dbl>	Syllable_of_object <dbl>
OSV	1.379310	7.123412
SOV	3.415796	4.239257

2 rows

As per Fig 3, the subject with focus highly prefers OSV order and when nothing is focused, SOV is preferred. In Tamil, suffixing the particles to subject and object is the most common way to focus them. Also, word order is not concerned when the elements are focused with particles and it can be either SOV or OSV. The word order deviates from SOV to OSV only when the subject is focused without any particles. But it is not necessary to prefer OSV in order to focus the subject. Because, it can be focused with the particles too.

Fig 3. Word order vs Focus



7. Conclusion

From these analyzes, it is clear that the length of subject, length of object and focus play a major role in deviating the word order. But the psych verbs are not having any influence towards the word order. Given these factors, the SOV is still higher than OSV. Even the focus with subject has a strong influence towards OSV, most of the time focus is taking place not through word order but through suffixing particles. In Tamil there is no rule or compulsion for the word order to deviate from their standard order i.e., SOV. Therefore, I am concluding my work by saying that the SOV word order is more prevalent with psych verbs of Tamil and moreover these verbs are not deviating the word order in Tamil language.

Bibliography

Arden, A. H. (1891). *A Progressive Grammar of Common Tamil*. Society for Promoting Christian Knowledge.

Belletti, A., & Rizzi, L. (1988). Psych-verbs and θ-theory. *Natural Language and Linguistic Theory*, *6*(3), 291–352. https://doi.org/10.1007/BF00133902

Download Corpora Tamil. (n.d.). Retrieved August 25, 2023, from https://wortschatz.uni-leipzig.de/en/download/Tamil#tam_newscrawl_2011

Landau, I. (2010). *The locative syntax of experiencers*. MIT Press.

Masloch, S., Poppek, J. M., Robrecht, A., & Kiss, T. (2021). *Syntactic Pattern Distribution Analysis of Experiencer-Object Psych Verbs*.

Murugaiyan, A. (2011, June). Mapping Language Change in Tamil: Corpus analysis and Computer Database Making. In *Conference papers, International Forum for Information Technology in Tamil, University of Pennsylvania, Philadelphia* (pp. 301-307).

Steever, S. B. (2019). *The Dravidian Languages*. Routledge.

Temme, A. (2019). *The peculiar nature of psych verbs and experiencer object structures*. https://doi.org/10.18452/19889

இராச திருமாவளவன். (n.d.). *வேற்றுமைத் தொகை | தமிழ் இணையக் கல்விக்கழகம் TAMIL VIRTUAL ACADEMY*. (n.d.). Retrieved August 22, 2023, from https://www.tamilvu.org/ta/courses-degree-a021-a0214-html-a0214211-6615

Appendix

A.1. 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.

2. **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)'.

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

   (1) *அங்கே அவர்கள் லோகாம்பிகையைக் கண்டு மகிழ்ந்தனர்.*

4. **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'.

(2) 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.

5. 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'.

(3) 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.

6. Phrasal_subject

This column indicates whether the subject contains more than one word or not. The values of the column are 'yes' and 'no'. In a sentence, the subject can be a single word: அரசர் (*arasar*-King) or pronoun: அவன் (*avan* - He) or more than one word: (4) is the subject with more than one word. When the subject contains more than one word, then the value is 'yes'. There is an exception to it. Sometimes, a single name has more than one word. For e.g. பால் தாக்கரே (Pal Thakare). In these cases, the value is 'no'. In addition, when the subject has only one word, then also the value is 'no'.
In Tamil when the temporal and locative occurs before the subject, then they belongs to the subject. Subject with temporal: சங்க காலத்தில்(temporal) மக்கள்(subject) (*sanga_kaalathil makkal* - people in sangam period) and subject with locative: மெரினாவில்(locative) பல்லாயிரக் கணக்கில் மக்கள்(subject) (*merinavil pallaayira kanakkil makkal* – thousands of people in Merina beach). These two examples are also considered as subjects.

(4) 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.

7. 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'.

(5)	***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**.

8. 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)'.

9. 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'.

10. 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'.

Owing to the fact that the adverbials are acting as the stimuli of subject for psych verbs, it was considered as one of the cases of object in this annotation though they are actually not objects according to linguistics.

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'.

(6)

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.

(7) *புலிக்கு* (*pulikku*, to a tiger / for a tiger)

(8)

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.

(9)

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.

(10)

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.

(11)

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**.

11. 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.

12. 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'.

13. 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'.

14. 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'.

(12)

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.

15. 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'.

16. 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.

17. 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.

18. Focus:

This column indicates which element is focused more in 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)'.

For focusing the elements, particles are highly used in Tamil. The particles like '*தான்* (*thaan*)', '*கூட*(*kooda*)' and '*உம்* (*um*)' are affixed to the subject or object inorder to emphasize them. It is to be noted that the subject can be emphasized by either particles or placing it before the verb. But the object is emphasized only through particles. Even when object lies before verb i.e., SOV, it doesn't get focused while the standard word order of Tamil is SOV.

When the order is OSV, then there are two possibilities. Either the subject is focused or the length of object is very big. When the object length is too big, it prefers first position in order to reduce the sentence complexity and to enhance better understanding. In this case, subject is not focused while the target is just to make the sentence clear, hence the value of this column will be 'none'. When the length of object is pretty normal or small and yet the order is OSV, it means the subject is focused. In this case, the value of this column is 'subject'.

If there are no emphasized particles and the order is SOV, then also the value is 'none'. Despite of 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)'.

19. 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.