Computational Challenges with Tamil Complex Predicates

Kengatharaiyer Sarveswaran
University of Moratuwa

Miriam Butt
University of Konstanz

Proceedings of the LFG'19 Conference

Australian National University

Miriam Butt, Tracy Holloway King, Ida Toivonen (Editors)

2019

CSLI Publications

pages 272–292

http://csli-publications.stanford.edu/LFG/2019

Keywords: complex predicates, FSM, Tamil, restriction operator, morphology-syntax interface

Sarveswaran, Kengatharaiyer, & Butt, Miriam. 2019. Computational Challenges with Tamil Complex Predicates. In Butt, Miriam, King, Tracy Holloway, & Toivonen, Ida (Eds.), *Proceedings of the LFG'19 Conference, Australian National University*, 272–292. Stanford, CA: CSLI Publications.

Abstract

This paper presents work in the context of the development of a computational ParGram style grammar for Tamil. The grammar is implemented via the XLE grammar development platform and contains a Finite-State Morphological analyser implemented via Foma. This paper reports on challenges for the implementation found with respect to V-V complex predicates in terms of the interaction with phonology (*Sandhi*) and the lexicon. In particular, we focused on the interaction of causation and passivisation with complex predication. This paper provides further evidence from Tamil complex predicates for the use of the Restriction Operator and also addresses issues with respect to complex predication at the morphology-syntax interface.

1 Introduction

This paper presents work in the context of the development of a computational ParGram (Butt et al. 1999) style grammar for Tamil.[1] The grammar is implemented via the XLE grammar development platform (Crouch et al. 2017) and contains a finite-state morphological (FSM) analyser implemented (Sarveswaran et al. 2019) via Foma (Hulden 2009). The work to date has mainly focused on the implementation of basic clause types and the inflectional morphology within the morphological analyser.

In pursuing this work, we encountered challenges with respect to the implementation of V-V complex predicates in terms of the interaction with phonology, the lexicon and derivational morphology. In this paper, we focus on the challenges arising with respect to the interaction of causation and passivisation within complex predicates. Similar but not identical issues have been noted for Turkish (Çetinoğlu 2009) and Urdu (Bögel et al. 2019), leading to the use of the Restriction Operator for passivisation, rather than the classical lexical rules of LFG. This paper provides further evidence for the use of the Restriction Operator from Tamil complex predicates and also addresses issues with respect to complex predication at the morphology-syntax interface that have not previously been encountered within ParGram.

Tamil is well known for its diverse types of V-V sequences (Steever 1987, 2005). Here we focus on an instance of V-V complex predication as discussed by Annamalai (2013). We illustrate how this type of complex predication is handled in the Tamil LFG grammar using the causative and passive constructions of two verbs: 'buy' and 'give', whereby 'give' functions as a light verb that adds a beneficiary to the overall predication. A particular challenge in Tamil is that the elements of complex predicates can either be found written together as a single word, or be separated into two tokens. However, phonological Sandhi phenomena apply irrespective of the expression

[1]We gratefully acknowledge funding from the DAAD (German Academic Exchange Office) in support of this research.

in terms of one or two tokens and are realised obligatorily within Tamil orthography. The phonological properties of one part of the complex predicate condition Sandhi rules on the other part, irrespective of whether these are written as one or two parts. While this points towards an overall realisation of one prosodic unit irrespective of the realisation in terms of one vs. two tokens, it poses a challenge for the computational implementation of morphology-syntax interface as the analysis of individual words within the morphological analyser must anticipate possible Sandhi rules triggered by complex predicate formation in the syntax. We show how this phenomena can be handled without an extension of the existing ParGram architecture.

2 Background

2.1 Tamil

Tamil is a Southern Dravidian language spoken natively by more than 80 million people across the world. It has been recognised as a classical language by the government of India since it has more than 2000 years of a continuous and unbroken literary tradition (Hart 2000). It is an official language of Sri Lanka and Singapore, and has regional official status in Tamil Nadu and Pondichchery, India.

Tamil words have been primarily divided into four types, namely: nouns, verbs, intensifiers/attributives, and particles in grammar books written by native grammarians (Thesikar 1957, Senavaraiyar 1938). However, more modern work provides a different type of classification (Nuhman 1999, Paramasivam 2011). Beyond the nature of their part-of-speech category, words in Tamil can be further classified into divisible and indivisible categories. A divisible word can have six parts, namely: root, suffix, medial particle, *chariyai*, *Sandhi* and alteration (Nuhman 1999, Senavaraiyar 1938), where medial particles can be tense markers, and *chariyai* is a phonological modifier which can be further divided into a euphonic marker and an oblique marker based on the function expressed by it (Lehmann 1993). The notion of *Sandhi* is elaborated upon in the next section. The alteration is a phonological change which is realised as such in the orthography.

(1)

வந்தனன் (vantanan)				
வா	த்(ந்)	த்	அன்	அன்
vaa	t(n)	t	an	an
root (வா-> வ)	*Sandhi* (த் -> ந்)	*medial*	*chariyai*	suffix

‘(He) came.’

Example (1) shows that how a divisible word can be sliced into different parts. However, not all the divisible words have all these six parts.[2] In (1), வா-> வ and த் -> ந் are called alterations.

2.2 சந்தி (Sandhi)

Internal *Sandhi* refers to a phonological process triggered across two morphs within a (prosodic) word. When such a process is applied at the boundary of two words it is referred to as external *Sandhi*. External Sandhi can occur when the second word begins with one of the following consonants: க் (k), ச்(c), த் (t), ப் (p). However, further licensing conditions also need to be met, as shown below. Internal *Sandhi* is purely morphophonological in nature, while external *Sandhi* is also subject to syntactic or semantic constraints. Example (2) shows an internal *Sandhi* [t], this is inserted because the past tense marker (t) follows a vowel. Since Tamil orthography closely reflects the phonology of the language, *Sandhi's* effects on the orthography must necessarily be dealt with by any Tamil computational grammar.

(2)

படித்தான் (padittaan)			
படி	-த்	-த்	-ஆன்
padi	-t	-t	-aan
study	-SAN	-PAST	-3SMR

'(He) studied.'

The examples in (3) and (4) illustrate a case of external *Sandhi*. The object ('bull') and the verb contain identical final (object) and initial (verb) phonological segments. However, in (3) the insertion of *Sandhi* [p] is obligatory: *Sandhi* must apply if there is an overt accusative on the object. However, as shown in (4), no *Sandhi* occurs when there is no accusative marker even though it is an equivalent construction in terms of segmental phonology, i.e. in both (3) and (4) /i/ is the final vowel in the noun preceding the verb பிடித்தான் (pidiththan).

(3)

கந்தன்	காளையைப்	பிடித்தான்
kanthan	kalai-yai-p	pidiththan
Kanthan.NOM	bull-ACC-SAN	catch.PAST.3SMR

'Kanthan caught the bull.'

[2]Abbreviations in the glosses are: VP=Verbal Participle; INF=Infinitive; 3SN=3rd Person Singular Neuter; 1S=1st Person, Singular; 3SMR=3rd Person, Singular, Masculine and Rational; PASS=Passive; SAN=Sandhi; RP= Relative Participle; IMP=Imperative; CAUS=Causative; NOM=Nominative; DAT=Dative; ACC=Accusative.

While having *Sandhi* in (3) is compulsory, including the *Sandhi* in (4) is considered ungrammatical. This thus illustrates that *Sandhi* is not conditioned by purely segmental phonological factors.

(4)

கந்தன்	ஒரு	காளை	பிடித்தான்
kanthan	oru	kalai	pidiththan
Kanthan.NOM	a	bull.NOM	catch.PAST.3SMR

'Kanthan caught a bull.'

The presence or absence of *Sandhi* is furthermore indicative of different underlying syntactic structures. For instance, as shown in (5)–(6), the *Sandhi* [c] surfaces when the token is an adjective in (5), but does not surface in (6), where a string identical item is functioning as a relative participle and is at the right edge of a relative clause boundary. In (5) the item is the adjective முக்கியம் (mukkiyam) 'important', in (6) it is a relative participle derived from the verb முக்கு (mukku) 'dip'.

(5)

அவன்	முக்கியச்	செய்திகளை	வாசித்தான்
avan	mukkuya-c	seithikal-ai	vasiththan
avan.NOM	important.ADJ-SAN	news.PL-ACC	read.PAST.3SMR

'He read the important news items.'

(6)

அவன்	முக்கிய	செய்திகளை	வாசித்தான்
avan	mukkuya	seithikal-ai	vasiththan
avan.NOM	dip.RP	news.PL-ACC	read.PAST.3SMR

'He read dipped news items.'

We assume that the presence or absence of *Sandhi* is related to whether items are phrased together prosodically or not, so that *Sandhi* occurs within a prosodic phrase, but not across prosodic boundaries (see also Lahiri & Fitzpatrick-Cole 1999 for Bengali). Dealing with such prosodically conditioned *Sandhi* provides a challenge for computational grammar development within the ParGram framework. In this paper we show how we can model the phenomena via an interaction of the FSM analyser with the computational syntax.

2.3 Verbs in Tamil

Verbal morphology on simplex verbs in Tamil expresses information about tense, mood, aspect, negation, interrogativity, emphasis, speaker perspective, sentience or rationality, and conditional and causal relations (Anna-

malai et al. 2014). The structure of a simple verb is <root> + <medial-particle> + <terminal-suffix>. However, there are cases where a euphonic particle is also added in the middle, in addition to the usual medial particle. The medial particle is mainly used to realise the tense of the verb. There are three values for tenses in Tamil: past, present and future (Pope 1979, Lehmann 1993, Paramasivam 2011).

The terminal-suffix of a finite verb is used to realise multiple types of information such as number, person, gender, and rationality (Pope 1979) or status (Lehmann 1993). As for other morphosyntactic features, Tamil has singular and plural as values for number, 1st/2nd/3rd person values, and three gender values, namely masculine, feminine and neuter. In addition to these three genders, epicene is also used as a fourth one to mark the 3rd person plural forms (Lehmann 1993). Entities in Tamil are fundamentally classified into rational or irrational. This split is based on the status of an entity: Entities are termed rational if they are perceived as being able to think on their own, whereas the rest are termed irrational. This is different from splits found otherwise in terms of human vs. non-human or animacy. For instance, infants are considered to be irrational like other animal or inanimate objects even though infants are human and animate. Further, when people behave as if they are insane, they are morphologically classified as irrational.

2.4 ParGram project

The Parallel Grammar (ParGram) Project (Butt et al. 1999, Butt & King 2002) aims to develop and implement large and wide coverage grammars for languages of different families. These parallel grammars are written collaboratively within the linguistic framework of LFG and with an agreed set of grammatical features by the project group members. The XLE (Xerox Linguistic Environment) (Crouch et al. 2017), which is a parsing and generation implementation of LFG provided by PARC, is used as a grammar development platform. In addition to putting effort into feature standardisation, the project also promotes similar analyses for similar phenomena across languages (Butt & King 2002), a property which is useful for crosslingual language applications like machine translation and information retrieval.

3 Complex Predicates

The study of complex predicates (CP) has received a great deal of attention in the linguistic literature, along with a number of distinct interpretations. We base this paper on the definition proposed by Butt (1995), which views CPs as being formed when two or more predicational units enter into a relationship of co-predication. Each predicational unit adds arguments to a mono-clausal predication; a similar definition or idea can also be found in

Mohanan (1994, 1997) and Alsina et al. (1997). In LFG, these two or more semantic heads correspond to a single PRED at the level of f-structure. C-structure does not determine CP status and the elements contributing a CP can be either morphological or syntactic (Butt 2010). However, regardless of whether the complex predication is morphological or syntactic, the composition of the arguments of both of the predicational units works according to the same principles (Alsina 1996).

Complex predicates are very common in Tamil (Annamalai 2013). For instance, verbs like வை (vay) 'place', விடு (vidu) 'let go', பார் (paar) 'see/look' may function as both main/full and light verbs. As light verbs, they mean 'cause', 'let' and 'try', respectively (Annamalai 2013).

3.1 Complex Predicates in Tamil

Tamil verbs have been analysed mostly from a prescriptive perspective, and most of these studies are based on the very first Tamil grammar called *Tholkapiyam*[3] and a derived piece of work, the *Nannool*, published in the 13th century CE. From the 18th century CE on Western scholars have also contributed to the study of Tamil grammar. However, except for the attempt by Annamalai (2013), no scholars have clearly articulated differences between complex predicates, serial verb constructions (Steever 2005, Fedson 1981), complex verbs (Agesthialingom 1971), and compound verbs (Agesthialingom 1971, Nuhman 1999, Fedson 1981, Paramasivam 2011). This stands in contrast to the work done for other South Asian languages like Urdu/Hindi (Butt 1995, Mohanan 1997, Butt & Lahiri 2013). However, it is important to understand the differences across these potentially confusing categories for the development of computational resources such as computational grammars (Butt & King 2002), WordNet (Chakrabarti et al. 2007), and machine translation (Kaplan & Wedekind 1993, Butt 1994).

As noted in the existing literature (Annamalai 2013, Steever 2005), Tamil is well known for diverse types of Verb-Verb (V-V) and Noun-Verb (N-V) constructions. Muthuchchanmugan (2005) shows that the main verb (also called lexical head word) in Tamil can be followed by up to four verbal units. However, whether all of them are auxiliaries as he claims, or not, is debatable. Tamil is a head-final language. In V-V constructions the terminal verbal unit is the final item in a sequence. The preceding verbal units can be in either an adverbial or infinitival form. The terminal verbal unit is the item that carries all the functional information such as tense, person, number, and gender. The V-V sequences are used to express a range of semantic information. This includes crosslinguistically well-established categories such as causative, passive, permissive, negation, aspectual information, and mood and modality, including obligation vs. possibility. The

[3]The date of publication is imprecise and uncertain, scholars argue that it could be between the 5th century BCE and the 5th century CE.

literature also describes definitive and conclusive meanings, the expression of irritation, carelessness, augmentation, prediction and intention (Paramasivam 2011, Muthuchchanmugan 2005). Tamil also has compound nominal predicates with N-N and V-N sequences, but these are not the topic of this paper. In what follows, we focus on the treatment of light verbs, causatives and passives in V-V constructions.

3.2 Light Verb Constructions

Light Verbs (LV) differ from main/full verbs in terms of their syntactic distribution and lexical semantics. While main verbs can stand alone and predicate independently, light verbs are dependent on the existence of another predicative element in the clause. LVs are light in the sense that they do not carry the meaning of the corresponding full verb, yet they still contain lexical semantic information (Butt 2010, Annamalai 2013). Unlike auxiliaries, they are not fully functional elements. Together with the main predicational element, in our case a verb, the light verb forms a syntactically monoclausal unit. Following (Butt 2010), we analyse LVs as a separate syntactic category and differentiate them from both main verbs and auxiliaries in the language. Such structures are common in South Asian Languages (SAL) (Butt & Lahiri 2013), including Tamil (Annamalai 2013).

Annamalai (2013) has analysed various V-V, Infinitive-V, N-V and Verbal Participle-V sequences and differentiates between Serial Verb Constructions (SVC) and Complex Predicates (CP). Example (7) illustrates a simple transitive verb வாங்கு (vangu) 'buy'. The same main verb used together with கொடு (kodu) 'give' in its light verb sense forms a CP in (8). The light verb 'give' contributes a beneficiary meaning to the predication and licenses the use of an additional beneficiary indirect object (OBJ-TH). The light verb as the terminal verbal unit carries the functional information, in this case with regard to tense, number and person.

(7)

நான்	காரை	வாங்கினேன்
naan	car-ai	vanginen
I.NOM	car-ACC	buy.PAST.1S

'I bought the car.'

(8)

நான்	அவனுக்குக்	காரை	வாங்கிக்கொடுத்தேன்
naan	avanukku-k	car-ai	vangikkoduththen
I.NOM	he.DAT-SAN	car-ACC	buy.VP-SAN.give.PAST.1SG

'I bought him a car.'

The f-structure in (9) shows an analysis of the complex predicate in (8) and illustrates the monoclausal nature of the co-predication. The f-structural analysis follows conventions established by the Urdu ParGram grammar (Butt & King 2007) and includes information about lexical semantics, which in this case only involves the information that the overall predicate is agentive. The co-predication is shown via the composed verbal PRED value, which brings together information contributed by each of the predicates.

(9)

$$\begin{bmatrix}
\text{PRED} & \text{`KODU} \left\langle (\uparrow \text{OBJ-TH}), \text{VANGU} \left\langle (\uparrow \text{SUBJ}),\ (\uparrow \text{OBJ}) \right\rangle \right\rangle\text{'} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`PRO'} \\ \text{PRON-FORM} & \text{NAAN} \\ \text{CASE} & \text{NOM} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 1 \end{bmatrix} \\
\text{OBJ-TH} & \begin{bmatrix} \text{PRED} & \text{`PRO'} \\ \text{PRON-FORM} & \text{AVAN} \\ \text{CASE} & \text{DAT} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`CAR'} \\ \text{CASE} & \text{ACC} \\ \text{DEF} & + \end{bmatrix} \\
\text{TNS-ASP} & \begin{bmatrix} \text{TENSE} & \text{PAST} \\ \text{MOOD} & \text{INDICATIVE} \end{bmatrix} \\
\text{LEX-SEM} & \begin{bmatrix} \text{AGENTIVE} & + \end{bmatrix} \\
\text{VTYPE} & \begin{bmatrix} \text{COMPLEX-PRED} & \text{VV} \end{bmatrix} \\
\text{STMT-TYPE} & \text{DECL} \\
\text{PASSIVE} & - \\
\end{bmatrix}$$

The example in (10) shows an alternative version of (8) in which the two parts of the complex predication are realised separately. The nature of the monoclausal co-predication at f-structure does not change with this alternative realisation. The *Sandhi* [k] is also triggered on the main verb வாங்கு (vangu) 'buy' just as in the single word realisation in (8). We discuss how this implementational challenge is resolved in section 4.2.

(10)

நான்	அவனுக்குக்	காரை	வாங்கிக்	கொடுத்தேன்
naan	avanukku-k	kar-ai	vangi-k	koduththen
I.NOM	he.DAT-SAN	car-ACC	buy.VP-SAN	give.PAST.1SG

'I bought car for him.'

3.3 Causatives

A causative in Tamil can be realised either morphologically or syntactically (Nuhman 1999, Paramasivam 2011). In either case, the causative is realised as a monoclausal complex predication (Annamalai 2013).

3.3.1 Morphological Realisation of Causatives

The causative can be realised in the morphology through three morphs: வி (vi), and (ப்)பி ((p)pi), which occur before the tense maker in a verb (Steever 2005). The choice of causative morph depends on the last vowel of the verb root and is thus phonologically conditioned.

The example in (11) shows how the morpheme *vi* is used in causatives. As shown in (12), the f-structure for (11) analyses the morphological causative as a CP (Butt 2010, Butt et al. 2003); the argument roles in this causative are causer.NOM and causee.INST; the case marking of the patient is NOM. The causative morpheme co-predicates together with the main verb.

(11)

வாங்குவித்தான் (vanguvittaan)

வாங்கு	-வி	-த்	-த்	-ஆன்
vangu	-vi	-t	-t	-aan
buy	-CAUS	-SAN	-PAST	-3SMR

'he made somebody buy (something)'

(12) $\left[\text{PRED} \quad \text{`CAUS}\left\langle (\uparrow \text{SUBJ}), \text{`VANGU}\left\langle (\uparrow \text{OBL-INST}),\ (\uparrow \text{OBJ})\right\rangle\text{'}\right\rangle\text{'}\right]$

The example in (13) shows that LV constructions and causatives can be stacked in the sense that the causative applies to the entire LV construction, irrespective of whether the two verbs are written together or separately. The corresponding f-structural analysis is shown in (14).

(13)

நான்	அவளைக்கொண்டு	அவனுக்கு	
naan	avalaikkodu	avanukku	
I.NOM	she.ACC.INST	he.DAT	

ஒரு	கார்	வாங்கிக்	கொடுப்பித்தேன்
oru	car	vangi-k	koduppitthen
a	car.NOM	buy.VP-SAN	give.CAUS.PAST.1SG

'I got her to buy a car for him.'

(14)

$$\begin{bmatrix}
\text{PRED} & \text{`CAUS}\left\langle (\uparrow \text{SUBJ}), \text{`KODU}\left\langle (\uparrow \text{OBJ-TH}), \text{`VANGU}\left\langle (\uparrow \text{OBL-INST}), (\uparrow \text{OBJ})\right\rangle'\right\rangle'\right\rangle' \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`PRO'} \\ \text{PRON-FORM} & \text{NAAN} \\ \text{CASE} & \text{NOM} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 1 \end{bmatrix} \\
\text{OBJ-TH} & \begin{bmatrix} \text{PRED} & \text{`PRO'} \\ \text{PRON-FORM} & \text{AVAN} \\ \text{CASE} & \text{DAT} \end{bmatrix} \\
\text{OBL-INST} & \begin{bmatrix} \text{PRED} & \text{`PRO'} \\ \text{PRON-FORM} & \text{AVAL} \\ \text{CASE} & \text{INST} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`CAR'} \\ \text{CASE} & \text{NOM} \\ \text{DEF} & - \end{bmatrix} \\
\text{TNS-ASP} & \begin{bmatrix} \text{TENSE} & \text{PAST} \\ \text{MOOD} & \text{INDICATIVE} \end{bmatrix} \\
\text{VTYPE} & \begin{bmatrix} \text{COMPLEX-PRED} & \text{VV} \end{bmatrix} \\
\text{PASSIVE} & - \\
\text{STMT-TYPE} & \text{DECL}
\end{bmatrix}$$

3.3.2 Syntactic Realisation of Causatives

Causatives in Tamil can also be realised syntactically by adding one of the following verbs after an infinitive form of the main verb: செய் (sei) 'do', வை (vai) 'put', பண்ணு (pannu) 'do'. These verbs do not predicate as full verbs in this case and have the character of light verbs, expressing the non-referential meaning 'make'. When one of these three verbs is used as the main verb of a predication, the other two verbs can function as causativising light verbs in the combination. As shown in (15), we again face the implementational challenge that the main verb and the causative light verb

can be written together as one token or separately as two tokens. Further, ப் (p) or ச் (c) will be added as a *Sandhi* to the main verb as part of the causativisation, for *pannu* 'do' and *sei* 'do', respectively. There is no *Sandhi* for *vai* 'put' as it begins with a consonant வ்(v).

(15)

அவனை	ஒரு	கார்	வாங்கச்	செய்தேன்
avan-ai	oru	car	vanga-c	seithen
he-ACC	a	car.NOM	buy.INF-SAN	make.PAST.3SM

'(I) made him buy a car.'

3.4 Passives

Passive constructions in Tamil are also realised via a V-V construction. The verb படு (padu) 'be touched/be experienced/sleep' is used in the passive constructions, where *padu* is an auxiliary verb. Together with an infinitive form of a main word, it gives the meaning of 'be subjected to' (Annamalai 2013).

For instance, consider (16) and its passive version in (17). As per the standard LFG lexical rule for passivisation, the original OBJ becomes the nominative SUBJ and the original SUBJ is realised as an instrumental adjunct (OBL-INST). The monoclausal analysis of the passive is shown below. As in causatives, CP passives can also be written as one word, for our example in (17) this is: வாங்கிக்கொடுக்கப்பட்டது (vangkikkodukkappaddathu). Passives can also be written as separate words as shown in (18). However, when 'give' is a light verb, according to the corpus analysis, the passive part is always written together with it as in (17).

(16)

ராம்	அவனுக்கு	ஒரு	கார்	வாங்கிக்	கொடுத்தான்
ram	avanukku	oru	car	vangki-k	koduththaan
ram.NOM	he.DAT	a	car.NOM	buy.VP-SAN	give.VP.PAST.3SG

'Ram bought a car for him.'

(17)

ராமால்	அவனுக்கு	ஒரு	கார்	வாங்கிக்	கொடுக்கப்பட்டது
ramaal	avanukku	oru	car	vangki-k	kodukkappaddathu
ram.INST	he.DAT	a	car.NOM	buy.VP-SAN	give.INF.SAN.PASS

'A car was bought for him by Ram.'

(18)

கொடுக்கப்	பட்டது
kodukkap	paddatu
give.INF.SAN	do.PAST.3SG
'was given'	

(19)

$$\begin{bmatrix}
\text{PRED} & \text{`KODU} \left\langle (\uparrow \text{OBJ-TH}), \text{VANGU} \left\langle (\uparrow \text{SUBJ}) \right\rangle \right\rangle\text{'} \\
\text{OBJ-TH} & \begin{bmatrix} \text{PRED} & \text{`PRO'} \\ \text{PRON-FORM} & \text{AVAN} \\ \text{CASE} & \text{DAT} \end{bmatrix} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`CAR'} \\ \text{CASE} & \text{NOM} \\ \text{DEF} & - \end{bmatrix} \\
\text{OBL-INST} & \begin{bmatrix} \text{PRED} & \text{`RAM'} \\ \text{CASE} & \text{INST} \end{bmatrix} \\
\text{TNS-ASP} & \begin{bmatrix} \text{TENSE} & \text{PAST} \\ \text{MOOD} & \text{INDICATIVE} \end{bmatrix} \\
\text{VTYPE} & \begin{bmatrix} \text{COMPLEX-PRED} & \text{VV} \end{bmatrix} \\
\text{STMT-TYPE} & \text{DECL} \\
\text{PASSIVE} & + \\
\end{bmatrix}$$

4 Grammar Implementation

4.1 Existing ParGram Strategies

A combination of the Restriction Operator (Kaplan & Wedekind 1993) and the CHECK feature have been used to handle complex predicates within ParGram, especially for causation and passivisation in languages that work similarly, but not identically to Tamil: Turkish (Çetinoğlu 2009) and Urdu (Bögel et al. 2019). We propose to use the ParGram framework and strategies to handle the Tamil V-V complex predication described in this paper and show how it can be extended to cover the morphonological challenges posed by external *Sandhi* phenomena.

The CHECK feature was introduced within ParGram as a way to handle well-formedness checking. Information that is only relevant for ensuring morphosyntactic well-formedness, but is not relevant for down-stream semantic interpretation or further "higher" Natural Language Processing applications can be stored here.

The Restriction Operator allows for the manipulation of f-structural information. Atrribute-feature values may be "restricted" out as shown in the f-structures in (20) and (21) where the CASE feature has been restricted out of (20) via an application of the Restriction Operator '/': (↑/CASE). This type of restricting out might become necessary if a value for CASE had already been assigned by one part of the grammar but needed to be changed by another part. This situation arises, for example, if a main verb with a certain lexical semantic specification enters into a complex predication and the lexical semantics of the main predication "change" under complex predicate composition.

(20)
$$\begin{bmatrix} \text{PRED} & \text{'RAM'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{CASE} & \text{NOM} \end{bmatrix}$$

(21)
$$\begin{bmatrix} \text{PRED} & \text{'RAM'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix}$$

Most importantly, however, the Restriction Operator provides the ParGram grammars with a means of combining PRED information as described in section 3 for the V-V complex predicates and causativisation in the absence of an implementation of Mapping or Linking Theory.[4] The use of the Restriction Operator for complex predicate formation (Butt et al. 2003, Butt & King 2003, Butt et al. 2008) and passivisation (Wedekind & Ørsnes 2003) has been described elsewhere and is not repeated here.

In the next section we show how we used these feature and design recommendations and extended them where necessary to develop the Tamil grammar. We have developed our own FSM analyser and generator (Sarveswaran et al. 2018, 2019) and this is integrated into our grammar development efforts and extended where necessary to meet the implementational challenges.

[4]There are no technical impediments to a implementation of Mapping or Linking Theory. However, in the absence of a theoretical consensus on the issue, the Restriction Operator has emerged as a mathematically well-defined way of dealing with predicate composition without affecting the underlyingly monotonic nature of the LFG implementation (Kaplan & Wedekind 1993). XLE operates with a rudimentary version of argument structure internally that bears some resemblance to Kibort's recent linking proposals (Kibort 2013, 2014, Crouch et al. 2017). This topic provides a potential for further fruitful exploration, but is out of the scope of this current paper.

4.2 Implementation and challenges

In this section, we discuss how we handle the interaction between *Sandhi*, complex predicates, causatives and passives in our Tamil grammar. We treat light verbs via the Restriction Operator in combination with morphophonological rules implemented in the FSM analyser ThamizhiFST (Sarveswaran et al. 2018, 2019). The CHECK feature was originally defined within ParGram to ensure of morphosyntactic wellformedness. Our treatment extends this feature to include the relevant *Sandhi* information.

4.2.1 Handing *Sandhi*

As shown in the previous sections, Tamil has two types of *Sandhi*, internal and external. Internal *Sandhi* can be handled entirely within our Tamil FSM ThamizhiFST. The treatment of internal *Sandhi* is fairly straightforward in that words with incorrect *Sandhi* patterns are not analysed and thus identified as misspellings.

However, external *Sandhi* has to be dealt with carefully. The morphological analyser is able to show whether a given word has a *Sandhi* letter at the end or not. However, it cannot check whether the *Sandhi* was used appropriately since that information will only become available as part of the syntactic analysis. We therefore extended the CHECK feature to check on the wellformedness of the morphophonology by ensuring that the correct *Sandhi* letter is indeed used.[5]

As shown in (22) for words with a *Sandhi* க் (k), we associate the f-structural information (↑CHECK _Sandhi-k) = + with the morphological tag provided by the FSM (see Kaplan et al. (2004) for details on the integration of an FSM with XLE). The morphological tag is part of the analysis of the Tamil word. Instances of Sandhi-p, Sandhi-c, and Sandhi-t are treated similarly.

(22) +Sandhi-k (↑CHECK _Sandhi-k) = +

This f-structural attribute is used to constrain the possible syntactic analyses in the grammar and to check whether the correct *Sandhi* has indeed been used in the syntactic context across two separate syntactic tokens.

[5]One reviewer suggested that *Sandhi* can be treated along the lines of initial mutations in Welsh as proposed by Mittendorf & Sadler (2006), through just the use of the morphological analyser, avoiding the need of a CHECK feature in the syntax. However, the Welsh initial mutation system does not follow from synchronically regular phonological rules, unlike in Tamil. In Tamil we also derive important syntactic information from the application or non-application of external *Sandhi* (e.g., adjectives vs. relative clauses). We therefore have decided to continue handling external *Sandhi* via a combination of FSM and the CHECK feature in the syntax.

4.2.2 Handling Complex Predicates in Tamil

When a CP is written as one token as in வாங்கிக்கொடுத்தேன் (vankikkodutteen) 'I bought for someone', we provide a subcategorisation frame as part of the lexical rule and handle it as a regular lexical item.

However, when a CP is written as two tokens as in (13), the two predicational units can be composed via the Restriction Operator. We follow the analyses established in Butt et al. (2003) and Butt & King (2006) whereby the light verb subcategorises for regular arguments as well as a variable %PRED2. This variable will be substituted in by the subcategorisation frame of the main verb as part of the complex predicate composition. For instance, in our grammar, we have a lexical entry for a light verb with its functional features கொடுத்தேன் (kodutteen) 'I gave', as shown in (23).

We then use the template in (24) which we obtained from Dalrymple et al. (2004) in the grammar rules for light verb to compose arguments.

(23)

கொடு Vlight XLE (↑ PRED)='கொடு<(↑OBJ-TH) %PRED2>' .

(24)

VV-ANNOTATION =

(↓ CHECK _RESTRICTED)	= +
(↑ PRED ARG2)	= (↓ PRED)
↓\PRED\SUBJ\CHECK\ OBJ-TH\OBJ\PASSIVE	= ↑\PRED\SUBJ\CHECK\ OBJ-TH\OBJ\PASSIVE
(↑ OBJ)	= (↓ OBJ)
(↑ SUBJ)	= (↓ SUBJ)
(↑ VTYPE COMPLEX-PRED)	= vv.

4.2.3 Handling Passives

Recall that the causative and passive are monoclausal constructions, but that the two predicational heads can be written either as one token or as separate tokens. If the two verbs are written together, they can be dealt with by the morphological analyser as shown in (25).[6] The surface form of the word is associated with tags shown in (25) via a series of rules. The analysis provides the information that this is a passive verb in the past and that it is combined with a 'give' light verb.

[6]Effects of assimilation and Sandhi are shown within parentheses.

(25)

வாங்கிக்கொடுக்கப்பட்டது
vangikkodukkappattathu
vang-i-(k)kodu-(k)k(ap)-pattatu
vangu +verb +vp +give +past +pass

The stem (vangu) and the tag for the light verb 'give' are straightforwardly associated with subcategorisation information via the stem lexicon contained in the grammar, as shown in (23) and (26).

(26) வாங்கு V XLE (↑ PRED)='வாங்கு<(↑ SUBJ)(↑ OBJ)>' .

Most of the attendant morphological tags are also straightforwardly associated with the corresponding f-structure information, e.g., (↑TNS-ASP) = PAST for the +past tag (Kaplan et al. 2004). However, an interesting question arises with regard to the treatment of the passive.

Classically, the passive is treated via lexical rules Bresnan (1982) in the ParGram grammars. These lexical rules are coded as part of the lexical entries, allowing for either an active or a passive version of the verb. The passivised version of vangu 'buy' in (27), for example, would result in the subcategorisation frame (↑ PRED) = 'vangu<(↑SUBJ)>' due to the application of the passive lexical rule.

(27)

vangu	(↑ PRED) = 'vangu<(↑SUBJ) (↑OBJ)>'
+past	(↑ TNS-ASP TENSE) = PAST
+3sg	(↑PERS) = 3
	(↑NUM) = sg

As outlined in section 3, passivisation can also be applied to a composed complex subcategorisation frame. For instance, the composed subcategorisation frame of the complex predicate வாங்கிக்கொடுத்தான் (vankikkoduthaan) '(he) bought' is 'give-buy <(↑OBJth) (↑SUBJ) (↑OBJ)>'. When it is passivised, the resulting subcategorisation frame would be 'give-buy <(↑OBJth) (↑SUBJ)>'. Passivisation via lexical rules is straightforwardly implementable when the parts of the verbal sequence are contained within one lexical item. In this case the lexical rule can be applied to the whole composed subcategorisation frame that is coming out of the lexicon.

However, an analysis via a passive lexical rule is not possible when the predicational heads are realised as two separate tokens. This is because the passive morpheme is morphologically attached to only one of the items in the verbal sequence and would naturally apply to only the subcategorisation frame of that item (Çetinoğlu 2009). However, passivisation actually needs to be applied to the composed subcategorisation frame of the complex predicate (which is distributed across two separate tokens in this case).

We therefore modified the existing lexical rule treatment of passivisation and instead developed an analysis in terms of the Restriction Operator, as shown in (28) (cf. also Wedekind & Ørsnes (2003)). The advantage of this analysis is that all operations or subcategorisation frames are now treated via the same mechanism and predicate composition can be treated in the same way irrespective of how the parts of a complex predication are realised: as two separate tokens or as a single complex verb, expressing the intuition that morphological and syntactic complex predication involves the same mechanism (Alsina & Joshi 1991).

5 Conclusion

This paper has examined the interaction between Tamil benefactive V-V complex predicates, causativisation, passivisation and attendant morpho-phonological *Sandhi* effects in the context of computational grammar development within LFG. Tamil orthography provides a particular challenge for grammar development as the verbal sequences can be optionally realised as one single token or several different tokens, but that the externaal *Sandhi* effects surface in either case. We showed how these external *Sandhi* effects can be dealt with via an interaction between the FSM and the grammar by utilising the concept of CHECK features introduced within ParGram and extending it to checking morphonological well-formedness. We further showed modeled the interaction of V-V benefactive complex predicates with causation via the Restriction Operator, but encountered problems when we attempted to handle the interaction with passivisation via classical lexical rules. We thus proposed to handle passivisation via the Restriction Operator as well, lending support to previous analyses along these lines for Urdu and Turkish (Çetinoğlu 2009, Butt et al. 2003) as well as Danish (Wedekind & Ørsnes 2003).

References

Agesthialingom, S. 1971. A Note on Tamil Verbs. *Anthropological Linguistics* 121–125.

Alsina, Alex. 1996. *The Role of Argument Structure in Grammar*. Stanford: CSLI Publications.

Alsina, Alex, Joan Bresnan & Peter Sells (eds.). 1997. *Complex predicates: Structure and theory*. Stanford: CSLI Publications.

Alsina, Alex & Smita Joshi. 1991. Parameters in causative constructions. In *Papers from the 27th Regional Meeting of the Chicago Linguistic Society*, 1–15.

Annamalai, E. 2013. The Variable Relation of Verbs in Sequence in Tamil.

NINJAL International Symposium 2013: Mysteries of Verb-Verb Complexes in Asian Languages.

Annamalai, E, A Dhamotharan & A Ramakrishnan. 2014. *Akarātiyiṉ putiya patippil taṟkālat tamiḻ ilakkaṇa viḷakkam* xxxi–xlvii. Crea-A Publishers, India.

Bögel, Tina, Miriam Butt & Tracy Holloway King. 2019. Urdu Morphology and Beyond: Why Grammars Should not live without Finite-State Methods. In Cleo Condoravdi & Tracy Holloway King (eds.), *Tokens of Meaning: Papers in Honor of Lauri Karttunen*, 439–465. Stanford: CSLI Publications.

Bresnan, Joan. 1982. The Passive in Lexical Theory. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 3–86. The MIT Press.

Butt, Miriam. 1994. Machine translation and complex predicates. In H. Trost (ed.), *Proceedings of the Conference KONVENS '94 "Verarbeitung natürlicher Sprache"*, 62–71. Springer, Berlin.

Butt, Miriam. 1995. *The Structure of Complex Predicates in Urdu*. Stanford: CSLI Publications.

Butt, Miriam. 2010. The Light Verb Jungle: Still Hacking Away. In M. Harvey M. Amberber & B. Baker (eds.), *Complex predicates in cross-linguistic perspective*, 48–78. Cambridge: Cambridge University Press.

Butt, Miriam & Tracy H. King. 2003. Grammar writing, testing and evaluation. In Ali Farghaly (ed.), *A Handbook for Language Engineers*, 129–179. Stanford: CSLI Publications.

Butt, Miriam, Tracy H. King & Gillian Ramchand. 2008. Complex predication: Who made the child pinch the elephant? In Linda Uyechi & Lian Hee Wee (eds.), *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, Stanford: CSLI Publications.

Butt, Miriam & Tracy Holloway King. 2002. Urdu and the Parallel Grammar project. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization, COLING*, 39–45. Association for Computational Linguistics.

Butt, Miriam & Tracy Holloway King. 2006. Restriction for Morphological Valency Alternations: The Urdu Causative. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, 235–258. CSLI Publications.

Butt, Miriam & Tracy Holloway King. 2007. Urdu in a Parallel Grammar Development Environment. *Language Resources and Evaluation: Special Issue on Asian Language Processing: State of the Art Resources and Processing* 41.

Butt, Miriam, Tracy Holloway King & John T. Maxwell III. 2003. Complex Predicates via Restriction. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG03 Conference*, 92–104. Stanford: CSLI Publications.

Butt, Miriam, Tracy Holloway King, Maria-Eugenia Nino & Frederique Segond. 1999. *A Grammar Writer's Cookbook*. Stanford: CSLI Publications.

Butt, Miriam & Aditi Lahiri. 2013. Diachronic pertinacity of light verbs. *Lingua* 135. 7–29.

Çetinoğlu, Özlem. 2009. *A large scale LFG grammar for Turkish*: Sabanci University dissertation.

Chakrabarti, Debasri, Vaijayanthi Sarma & Pushpak Bhattacharyya. 2007. Complex predicates in Indian Language Wordnets. *Lexical Resources and Evaluation Journal* 40(3-4).

Crouch, Richard, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III & Paula Newman. 2017. XLE documentation. Palo Alto Research Center. http://www.parc.com/istl/groups/nltt/xle/xle_toc.html.

Dalrymple, Mary, Ronald M Kaplan, Tracy Holloway King et al. 2004. Linguistic generalizations over descriptions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG04 Conference*, 199–208. Stanford: CSLI Publications.

Fedson, Vijayarani Jotimuttu. 1981. *The Tamil serial or compound verb*: University of Chicago, Department of Linguistics dissertation.

Hart, George L. 2000. Statement on the Status of Tamil as a Classical Language. https://southasia.berkeley.edu/tamil-classes.

Hulden, Mans. 2009. FOMA: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 29–32. Association for Computational Linguistics.

Kaplan, Ronald M., John T. Maxwell III, Tracy Holloway King & Richard Crouch. 2004. Integrating Finite-state Technology with Deep LFG Grammars. In *Proceedings of the European Summer School on Logic, Language and Information (ESSLLI) Workshop on Combining Shallow and Deep Processing for NLP*, .

Kaplan, Ronald M. & Jürgen Wedekind. 1993. Restriction and correspondence-based translation. In *Proceedings of the 6th Conference of the Association for Computational Linguistics European Chapter (EACL)*, 193–202. Utrecht University.

Kibort, Anna. 2013. Objects and Lexical Mapping Theory. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG13 Conference*, Stanford: CSLI Publications.

Kibort, Anna. 2014. Mapping out a construction inventory with (Lexical) Mapping Theory. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG14 Conference*, 262–282. Stanford: CSLI Publications.

Lahiri, Aditi & Jennifer Fitzpatrick-Cole. 1999. Emphatic Clitics and Focus Intonation in Bengali. In R. Kager & W. Zonneveld (eds.), *Phrasal*

phonology, Dordrecht: Foris Publications.
Lehmann, Thomas. 1993. *A grammar of modern Tamil*. Pondicherry Institute of Linguistics and Culture, India.
Mittendorf, Ingo & Louisa Sadler. 2006. A Treatment of Welsh Initial Mutation. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG06 Conference*, 343–364. Stanford: CSLI Publications.
Mohanan, Tara. 1994. *Argument Structure in Hindi*. Stanford: CSLI Publications.
Mohanan, Tara. 1997. *Multidimensionality of representation: NV complex predicates in Hindi*. Stanford: CSLI Publications.
Muthuchchanmugan. 2005. *Ikkala Mozhiyiyal (Modern Linguistics)*. Mahin Press, Chennai, India.
Nuhman, M.A. 1999. *Basic Tamil Grammar (In Tamil)*. Sri Lanka: Readers' Association.
Paramasivam, K. 2011. *Contemporary Tamil Grammar*. India: Adaiyaalam.
Pope, George Uglow. 1979. *A handbook of the Tamil language*. Asian Educational Services.
Sarveswaran, K, Gihan Dias & Miriam Butt. 2018. ThamizhiFST: A Morphological Analyser and Generator for Tamil Verbs. In *2018 3rd International Conference on Information Technology Research (ICITR)*, 1–6. IEEE.
Sarveswaran, K, Gihan Dias & Miriam Butt. 2019. Using meta-morph rules to develop morphological analysers: A case study concerning Tamil. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, 76–86. Dresden, Germany: Association for Computational Linguistics.
Senavaraiyar. 1938. *Tholkappiyam - Eluththathikaram*. Chunnakam, Sri Lanka: Thirumahal Press.
Steever, Sanford B. 1987. *The Serial Verb Formation in the Dravidian Languages*. Delhi: Motilal Banarsidass.
Steever, Sanford B. 2005. *The Tamil Auxiliary System*. New York: Routledge.
Thesikar, Sangara Namashivaya. 1957. *Nannool Viruthiyurai*. Chennai, India: Vithiyanubalana Press.
Wedekind, Jürgen & Bjarne Ørsnes. 2003. Restriction and Verbal Complexes in LFG: A Case Study for Danish. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG03 Conference*, 424–450. Stanford: CSLI Publications.