Mixed Categories in Tamil via Complex Categories

Miriam Butt
University of Konstanz

S. Rajamathangi
Jawaharlal Nehru University

Kengatharaiyer Sarveswaran
University of Moratuwa

Proceedings of the LFG'20 Conference

On-Line

Miriam Butt, Ida Toivonen (Editors)

2020

CSLI Publications

pages 68–88

http://csli-publications.stanford.edu/LFG/2020

Keywords: Mixed Category, Complex Category, Nominalization, Complementizer, relative clause, Tamil

Butt, Miriam, Rajamathangi, S., & Sarveswaran, Kengatharaiyer. 2020. Mixed Categories in Tamil via Complex Categories. In Butt, Miriam, & Toivonen, Ida (Eds.), *Proceedings of the LFG'20 Conference, On-Line*, 68–88. Stanford, CA: CSLI Publications.

Abstract

This paper discusses nominalized complements in Tamil, analyzing them as a type of mixed category. We unpack the complex morphological marking found on the nominalized complements and analyze their morphosyntactic properties. The embedded clauses function as verbally headed complements internally, but display nominal properties with respect to the matrix clause. We tie our analysis to a diachronic perspective on mixed categories and propose that the concept of complex categories developed within ParGram allows for: 1) an elegant account of the mixed categorical properties of Tamil nominalized complements; 2) factoring in the gradual effects of historical reanalysis.

1 Introduction

In this paper we discuss Tamil constructions as illustrated in (1) and (2), where the nature of the embedded complement is interesting. In (1) the embedded clause functions as COMP, but is morphologically a nominalized version of a relative clause. In (2) we have a nominalized version of a relative clause formed on top of a complementizer which is historically derived from the verb *en* 'say'. The examples show two seemingly contradictory features.[1]

(1) [avan pizhai sey-t-a-athu-ai] ram
[he mistake do-PAST-REL-PRON.3SN-ACC] Ram.NOM
nirupi-tt-aan
prove-PAST-3SM
'Ram proved (it) that he made mistakes.'

(2) [avan pizhai sey-tt-aan enp-a-athu-ai] ram
[he mistake do-PAST-3SM COMP-REL-PRON.3SN-ACC] Ram.NOM
nirupi-tt-aan
prove-PAST-3SM
'Ram proved (the fact) that he made mistakes.'

[†]We thank the DAAD (German Academic Exchange Office) for funding an International Summer School on Natural Language Engineering (ISSALE) in Colombo, Sri Lanka. This served to introduce the three authors to one another. We also thank the DAAD for funding that allowed K Sarveswaran to spend an extended time at the University of Konstanz via a personnel (PPP) exchange program that additionally supports the creation of Natural Language Processing (NLP) resources for Tamil. S Rajamathangi was also able to spend some time at the University of Konstanz via funding from the DFG, the German National Science foundation. This paper is a direct result of being able to come together to talk about Tamil NLP and Tamil syntax. Finally, we thank our two reviewers for helping to improve this paper considerably.

[1]Abbreviations are as follows: COMP=complementizer, NOM=nominative, ACC=accusative, GEN=genitive, PRON=pronoun, REL=relativizer, 1S=first person singular, 3SM=third person singular masculine, 3SN=third person singular neuter, 3PL=third person plural, FUT=future, COND=conditional, NEG=negation, PTCP=participle, PASS=passive, NOMZ=nominalizer, PERF=perfective, QUOT=quotative.

For one, as Amritavalli and Jayaseelan point out, "we have the embarrassment of tense inside gerunds" (Amritavalli and Jayaseelan 2008, §3.2). For another, the embedded subject within this nominalized embedded clause is nominative (verbally licensed), rather than genitive (nominally licensed) as we would expect from Stowell's (1981) Case Resistance Principle. Finally, in both examples the embedded clause is marked accusative.

We see these examples as instances of a type of mixed category and propose to analyze them via formal machinery first introduced in the computational ParGram[2] context, namely *complex categories*.

The next section provides some general background, section 3 presents the Tamil data. Section 4 discusses previous approaches to mixed categories within LFG and introduces the formal machinery of complex categories. Section 5 provides our complex category analysis and section 6 concludes.

2 Background and Motivation

Tamil is a Southern Dravidian language spoken natively by more than 80 million people across the world. It is recognized as a classical language by the Indian government due to over 2000 years of a continuous literary tradition.[3] It is an official language of Sri Lanka and Singapore, and has regional official status in Tamil Nadu and Pondichchery, India.

Despite its large speaker population and historical time depth, Tamil is an under-researched language that is also under-resourced from the perspective of Natural Language Processing (NLP). As part of a collaborative effort we have been working on creating resources for Tamil NLP by building a ParGram style (Butt et al. 1999) Tamil grammar, which includes a morphological analyzer. The grammar is implemented with the XLE development platform (Crouch et al. 2017), the morphological analyzer (Sarveswaran et al. 2019, 2018) is realized in FOMA (Hulden 2009).

One of our goals is to build a treebank for Tamil by using the Tamil ParGram grammar. To this end, we are using Tamil educational textbooks as our corpus and are also adding to the existing parallel ParGramBank (Sulger et al. 2013) on the INESS site (Rosén et al. 2012).[4] In going through our body of examples, we encountered a number of challenging phenomena, one of which we tackle in this paper, namely, nominalized complements.

3 Tamil Nominalized Complements

The morphological structure of the complements in examples (1) and (2) is complex. Both examples employ a relativization strategy to accomplish com-

[2]https://pargram.w.uib.no

[3]https://southasia.berkeley.edu/tamil-studies-initiative

[4]http://clarino.uib.no/iness

plement embedding, a process which is found in Dravidian more generally. One way to form nominal complements in Tamil involves the relativization of the embedded verb (1). Another strategy is to mark the complementizer with relativizing morphology (2).

3.1 Complementizers in Tamil

Tamil does not have complementizers of the *that*-type as in English. Rather, it uses a grammaticalized form of the verb *en* 'say'. In the examples below this is the frozen past participle form *enṛu*, which has been analyzed as a type of quotative (Amritavalli 2013, Balusu 2020). This is illustrated by (3), which is ambiguous between a quotative use and a complementizer reading. (4) illustrates a purely complementizer reading. Note that matrix complementation verbs can also take an accusative object that serves as a type of co-referent for the complementizer clause (4-b). In this case, we have a relativized structure, marked by the relative marker *-a*. Note that the resulting form is *enṛa* due to phonological processes.

(3) ravi [naan en nanban-ai santhi-tt-en] enṛu
Ravi.3SM.NOM [Pron.1S my friend-ACC meet-PAST-1S] QUOT
so-nn-an
say-PAST-3SM
'Ravi said that — "I saw my friend".'
'Ravi said that I saw my friend.'

(4) a. ravi [mazhai var-um enṛu] ninai-tt-aan
Ravi.3SM.NOM rain come-FUT.3SN COMP think-PAST-3SM
'Ravi thought that it will rain.'

b. [avan pizhai sey-tt-aan enṛ-a] **unmaiy-ai** ram
he mistake do-PAST-3SM COMP-REL truth-ACC Ram.NOM
nirupi-tt-aan
prove-PAST-3SM
'Ram proved the truth that he made mistakes.'

While the original meaning of *en* as 'say' remains transparent to speakers of Tamil, it is no longer in general use as a verb of communication. The Tamil situation is consistent with grammaticalization processes found in other languages. For instance, Klamer (2000) shows how verbs of reporting in Austronesian languages become quotatives and from there begin to function as complementizers.

Recall that Tamil has a long diachronic record. However, this diachronic information is difficult to access because Tamil is severely under-researched. Conducting an in-depth diachronic investigation goes beyond the scope of this paper, but a quotative use of the form *enṛu* can be found as far back as 450–500 CE (dates according to Zvelebil 1974), see (5).

(5) ira-pp-an ira-pp-aar-ai ellaam ira-pp-in kara-pp-aar
beg-FUT-1S beg-FUT-3PL-ACC all beg-FUT-COND hide-FUT-3PL
ira-van-min **enṛu**
beg-NEG-3PL QUOT
I will beg from all beggars "If you want to beg, do not beg from people who hide things they have." (*Kural*-1067, *Thirukkural*, 450–500 CE)

3.2 Relative Clauses in Tamil

Relative clauses (RCs) in Tamil do not have relative pronouns like in English. RCs are formed by adding an *-a* morpheme to a verb (6-a). In the future participle form with *-um*, the relative marker is null, as shown in (6-b). Krishnamurti (2003) analyzes the *-a* in RCs as an adjectivizing morpheme and the resulting "relative participles" as having an originally adjectival structure. We take no position on this analysis. In what follows we refer to the morpheme *-a* as a relativizer.

(6) a. [angu **nin-ṛ-a**] paiyan-ai naan paar-t-en
there **stand**-PAST-REL boy-ACC I.NOM.1S see-PAST-1S
'I saw the boy who stood there.'
b. [angu **nirk-um-∅**] paiyan-ai naan paar-pp-en
there **stand**-FUT-REL boy-ACC I.NOM.1S see-FUT-1S
'I will see the boy who will stand there.'

The head noun of the RC in (6) is 'boy'. But, in predicative contexts, one also finds RCs without a head noun, as in (7). In this case, the verb in the RC instead carries a pronominal form *-athu*. This *-athu* is form-identical with the indefinite pronoun *athu*.

(7) [angu nin-ṛ-a-**athu**] en thambi
there stand-PAST-REL-PRON.3SN my brother
'The **one** who stood there is my brother.'

In order to account for this, we posit a process of cliticization of the matrix clause pronoun onto the RC so that the pronoun is prosodically incorporated into the RC, with (8) showing a synchronically unattested unincorporated version we postulate as the source construction. This is in line with the general tendency of function words to cliticize (e.g., Selkirk (1995); for pronouns in particular see Lahiri et al. (1990), Bögel (2015)).

(8) [angu nin-ṛ-a] athu en thambi

The cliticization also took place in non-predicative contexts. The example in (9-a) involves a full head noun 'boy' in the accusative as the matrix object. In (9-b) an accusative pronoun *-avan* 'he' is substituted in. The head noun is outside of the RC, the pronoun is realized as part of the RC.

(9) a. [angu nin-ṛ-a] **paiyan-ai** naan paar-t-en
there stand-PAST-REL boy-ACC I.NOM.1S see-PAST-1S
'I saw the boy who stood there.'

b. [angu nin-ṛ-a-**van**-ai] naan paar-t-en
there stand-PAST-REL-PRON.3SM-ACC I.NOM.1S see-PAST-1S
'I saw the one (he) stood there.'

Having looked at relativization strategies in Tamil, we are now ready to unpack our introductory examples.

3.3 Nominalized Relative Clause

We begin with the nominalized relative, repeated in (10) from (1). We can now identify the indefinite pronoun *athu* 'one' within the complement, as well as the relativizer *-a*. Following the general pattern found with RCs, the relativizer is attached to a participle form of the embedded verb.

(10) [avan pizhai sey-t-a-athu-ai] ram
[he mistake do-PAST-REL-PRON.3SN-ACC] Ram.NOM
nirupi-tt-aan
prove-PAST-3SM
'Ram proved (it) that he made mistakes.'

Also in analogy with the pattern found with RCs, the accusative *athu-ai* 'one' has been prosodically attached to the relativized verb, with the source construction having an NP outside of the COMP, the possibility of which was illustrated in (9-a). The *athu* 'one' thus functions as the matrix object and as such is marked accusative.

Overall, we therefore have a structure that is originally an RC meaning something like: 'Ram proved it, that he made a mistake.' This type of modification of an indefinite head pronoun is very close to a complementizer reading and we posit that this is what results.

Although we hypothesize that the attachment of the *athu-ai* 'one-ACC' is the result of prosodic incorporation, we have no synchronic evidence for clitic status. Rather, the forms are unequivocally treated as affixes in the literature (Rajendran 2012, Lehmann 1993, Krishnamurti 2003) so that the structures must now be analyzed as mixed categories which have an "outer" nominal structure built on a relativized clause that has an "inner" verbal structure, except that because the embedded verb is in a participle form, there is no subject-verb agreement with the embedded subject. We find no complementizer as such in this construction. Rather, the relativization of the embedded verb provides the function and meaning of complementation.

3.4 Nominalized Complement

We are now ready to analyze the nominalized complement, repeated here in (11) from (2). The indefinite neutral pronoun *athu* 'one' is again found in the embedded clause, as well as the relativizer *-a*.

(11) [avan pizhai sey-tt-aan enp-a-athu-ai] ram
[he mistake do-PAST-3SM COMP-REL-PRON.3SN-ACC] Ram.NOM
nirupi-tt-aan
prove-PAST-3SM
'Ram proved (the fact) that he made mistakes.'

We posit that in analogy to the general pattern found with RCs, the accusative *athu* 'one' has been prosodically attached to the complementizer, with an original structure having had an NP outside of the COMP, the possibility of which has already been illustrated by (4-b).

The *enp* in (11) is a form of the verb *en* 'say', the verb we discussed as undergoing reanalysis as a complementizer. The form *enṛu* is a frozen past participle form and functions most like a "pure" complementizer. However, the underlying verb *en* 'say' has several other participle forms and can appear with the relativizer (*-a*/∅) in all of these forms: *enṛ-a* (past), *enkiṛ-a* (present) and *enum-*∅ (future).

The forms *enṛathu*, *enkiṛathu* and *enpathu (enp-a-athu)* are essentially nominalized versions where the third person singular neuter pronoun *-athu* has been incorporated on top of the relative marker as in (11). Thus, if we unpack the complementizer form, we have a participle form of the verb *en* 'say', followed by the relativizer *-a*, followed by a form that was originally a pronoun *-athu*, which is in the accusative case *-ai*.

The overall original structure giving rise to these nominalized complements again parallels that of RCs. The difference between examples such as in (11) and what we have called a nominalized relative in (10) is the presence of the complementizer/quotative (cf. section 3.1) within the embedded clause. The accusative marking is a result of *-athu* originally being treated as a complement of the matrix verb, cf. (4-b).

The presence of the quotative/complementizer within the embedded clause has both syntactic and semantic effects on the embedded complement. In terms of syntax, the embedded verb in (11) anchors tense and shows subject-verb agreement, unlike the nominalized relative in (10). In both structures the embedded verbs predicate fully.

In terms of semantics, the presence of the quotative/complementizer appears to make an interpretational difference. As first reported by Lehmann (1993), nominalized complements as in (11) embed factive complements. That is, the embedded clause must represent a true proposition.[5]

[5] A reviewer notes that evidentiality is likely to play a role. We agree that this needs

Note that while we have identified the individual parts of the nominalized complementizer, the existing literature treats items like *athu* as pronominal suffixes with a nominalizing function (Krishnamurti 2003, Lehmann 1993).

3.5 Nominalizations in Tamil

Rajendran (2001) distinguishes between several different kinds of nominalizations in Tamil. One category involves nominalizing suffixes which are added directly onto the verb root, as illustrated in (12) with the suffix *-tal*. The second category involves nominalization of adjectival participial forms as in (1), the third the nominalization through complementizers as in (2).

(12) ram [kumar-in pizhai sey-tal-ai]
Ram.NOM.3SM Kumar.3SM-GEN mistake do-NOMLZ-ACC
so-nn-aan
tell-PAST-3SM
'Ram told (of) Kumar's doing wrong.'

Like in our running examples (1) and (2), the nominalized clause functions as the object of the matrix clause and is appropriately marked with the accusative case. In contrast to our running examples, however, the agent argument of the embedded verb is nominally licensed and is therefore realized with the genitive case. These verbal nouns are a classic case of mixed categories as they show both verbal and nominal properties. The arguments of the embedded clause are inherited from the verbal base, but the agent cannot be verbally licensed. Rajendran (2001) accounts for the differences between examples such as (12) and our running examples by positing nominalization at the sentence ((1) and (2)) vs. the lexical (12) level. Rajendran (2001) further notes that the nominalized complements and relatives are modifiable by adverbs, not adjectives, indicating an internal verbal structure. On the other hand, while the nominalized complements and relatives can be case marked, they cannot receive inflectional plural morphology. This indicates a less than full alignment with overall nominal properties.

Schiffman (1969) and Arden (1962) use a slightly different categorization and nomenclature in their studies of Tamil nominalizations, but both include (1) and (2) as instances of morphological nominalization.

Before moving on to our own analysis of complement nominalizations, we briefly touch on the issue of scrambling. Tamil allows scrambling of its major constituents in a clause, but generally shows restrictions within NPs. A natural question to inquire into is the scrambling possibilities of the various nominalized structures. We find that the nominalized relative (1), the nominalized complementizer (2) and the verbal noun (12) do not differ in terms of scrambling: all allow scrambling of all major constituents

to be investigated more deeply.

within the embedded clause, but the nominalized verb or complementizer is generally the final element in the embedded clause. This is as expected if the embedded clause is headed by a verb.

4 Mixed Categories and Complex Categories

We propose to analyze the complementizers found in our core examples in (1) and (2) as instances of **mixed categories**. Internally to the complementizer clause we have a verb (*sey*, V) and complementizer (*en*, C), respectively. However, the V and the C carry relativizing and nominalizing morphology. As discussed above, the current complementizer strategy appears to have evolved through a combination of diachronic developments within Dravidian. This fits with historical change having been identified as one reason for the existence of mixed categories (Nikitina 2008): One category is reanalyzed as another and gradually accumulates more of the properties associated with the "new" category during the change. Our Tamil complements seem to be classic examples of change in progress in that a verb of communication ('say') is being reanalyzed as a complementizer via an intermediate stage as a reportative/quotative (cf. Klamer 2000). Indeed, native speakers perceive the combination of relativizer+pronoun+case as an unanalyzable unit, indicating that language change is taking place.

In what follows, we first briefly discuss previous analyses of mixed categories within LFG, then introduce the formal notion of **complex categories** as implemented within the XLE grammar development platform (Crouch et al. 2017). In section 5 we then show how we propose to use this formal mechanism to model the phenomena associated with mixed categories.

4.1 Mixed Categories in LFG

The literature on mixed categories is large, with several different approaches having been put forward. A central problem posed by mixed categories is how to characterize them. One could simply admit categories such as VN (nominalizations) or VA (deverbal adjective) to one's inventory of categories, but the question then arises as to what the full inventory of categories should be and whether it is language universal. Computational efforts at defining inventories for Part-of-Speech (POS) tagging (Jurafsky and Martin 2009) have differed considerably, with tag sets having been proposed that range from including less than 20 POS tags to over a hundred. The Universal POS tag set arrived at by Universal Dependencies effort posits 14 basic word classes, none of which include mixed categories.[6] The reason for this perhaps is that mixed categories tend to be the result of the application

[6] https://universaldependencies.org/u/pos/

of derivational morphology: it seems counterintuitive to include categories derived by morphological processes in a basic inventory.

A different approach is represented by a definition of syntactic categories through feature bundles. A classic and simple approach involves the feature set [±N, ±V] (Chomsky 1981). Within LFG the feature set [±predicative,±transitive] has been used (Bresnan 2001, Bresnan et al. 2016). More complex feature bundles seek to model relevant morphological, syntactic and semantic properties, other approaches work with notions of prototypicality (Croft 1991) or canonical categories (Corbett 2006, 2007). Each of these proposals comes with their own set of resulting challenges and shortcomings. See Nikolaeva and Spencer (2020) and Lowe (2016) for comprehensive overviews and discussion.

Nikolaeva and Spencer (2020) develop an HPSG-inspired approach to adjectivized nouns that are able to modify other nouns. As part of their discussion, they define several different types of mixed categories. Our Tamil constructions fit the definition of *syntagmatic mixing*, by which a derived form displays distributional and selectional properties from the underlying category as well as the derived category (Nikolaeva and Spencer 2020, 24).

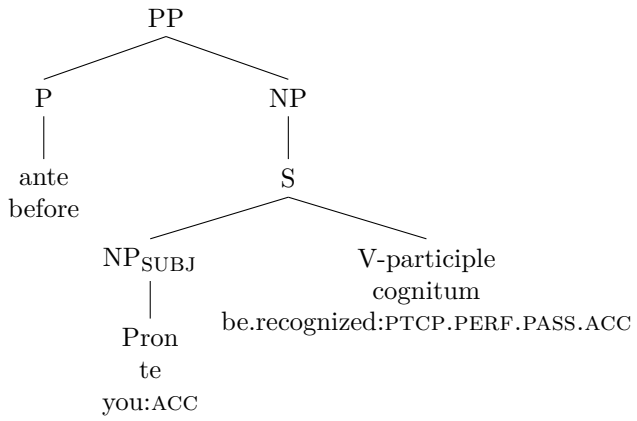
This syntagmatic mixing presents a problem for formal syntactic analyses that operate with principles governing the projection of words into phrases. Since the excesses of Transformational Grammar, formal syntax has developed an understanding that categories should not project randomly into phrases (e.g., so that an adjective heads a CP or a noun an IP), but be governed by constraints, such as X′ syntax (Bresnan 1976). Within LFG, the central problem posed by syntagmatic mixing can be stated in terms of the Principle of Endocentricity, which expects that "every phrasal projection has a unique lexical head which determines its categorial properties" (Bresnan and Mugane 2006, 203).

Work within LFG has offered up several different approaches to solve this fundamental violation of endocentricity. Central among these is the application of the theory of extended heads (Bresnan et al. 2016) by which lexical (but not functional) categories are assumed to have an extended head. This extended head mostly works out to be a functional category such as I or D. Bresnan et al. (2016) illustrate this analysis with respect to English gerunds and Bresnan and Mugane (2006) apply it to explain the properties of agentive nominalization in Gīkūyū. Nikitina and Haug (2016) appeal to the English gerund analysis by Bresnan et al. (2016) and propose a parallel analysis of Latin 'dominant participles'. LFG's projection architecture very naturally allows for more than one c-structure node to project to the same f-structure, and the extended head theory governs which types of c-structure nodes may serve to co-predicate, thus constraining the range of c-structural possibilities while accounting for mixed categorical properties.

The c-structure in (13) shows how Nikitina and Haug (2016, 38) treat Latin deverbal participles, which are analyzed as instances of clausal nomi-

nalizations. The verbal properties are licensed by the V within an exocentric category S, the nominal properties by the NP dominating the S.

(13)



Although agentive nominalizations in Gīkūyū work very differently from our Tamil complementizers and the Latin dominant participles, the analysis from Bresnan and Mugane (2006, 230) serves to illustrate how the projection across different nodes in the c-structure works. The lexical entry for the nominalized form in (14) contains a subcategorization frame that is licensed by the underlying verb. The lexical entry also contains information which ensures that the word must be part of both a nominal and a verbal projection. As the f- and c-structure in Figure 1 show, this is indeed ensured, with the N, NP and VP nodes all contributing to the same f-structure, thus accounting for the mixed properties of the agentive nominalization.

(14) mũthĩĩnji: N: (↑PRED) = 'slaughterer<<(↑OBJ)>$_v$>$_n$'
v: VP ε Cat(ϕ^{-1}(PRED ↑))
n: NP ε Cat(ϕ^{-1}(PRED ↑))

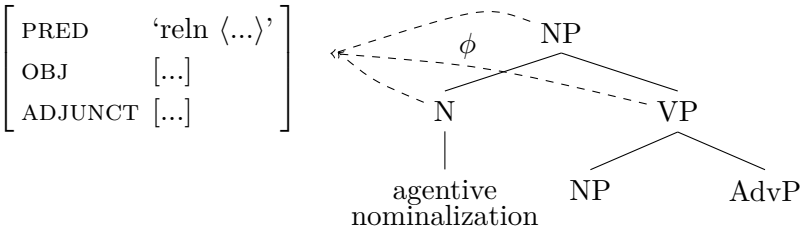


Figure 1: Analysis of Gīkūyū Agentive Nominalizations

While the analyses of Latin dominant participles and Gīkūyū agentive nominalizations provide crucial insights into their behavior and structure, the lexical entry in (14) taken together with the tree in Figure 1 means

that an unheaded VP is postulated in order to satisfy the mixed category requirement. And while Nikitina and Haug (2016) appeal to the analysis of English gerunds by Bresnan et al. (2016), it is not clear how the analysis of Latin participles conforms to the extended head principle, since the deverbal participle projects to an exocentric S. Furthermore, given that we have no independent evidence for a DP in Tamil and we have a situation in which a V in principle projects to a CP which in turn projects to an NP, it is not clear to us how we could straightforwardly apply an extended head analysis.

Nikolaeva and Spencer (2020) put forward a very different and complex proposal that focuses on modeling the lexical relatedness between basic and derived forms through an interplay between morphology, syntax and lexical semantics. We particularly find the argument-structure considerations introduced by Spencer (2015) and Nikolaeva and Spencer (2020) important, but these are less relevant for our Tamil complementizers. Indeed, Lowe (2016) takes stock of the literature on mixed categories and argues that phenomena which involve a consistent internal syntax coupled with a consistent external distribution are not true instances of mixed categories. He also suggests that these "lesser" versions of mixed categories could be treated by the formal means of complex categories, as developed within XLE (Crouch et al. 2017). Our Tamil nominalized complement structures mostly display a consistent internal syntax (C/V) and a consistent external distribution (N), with some differences being the inability to take plural morphology despite the external N distribution and the absence of subject-verb agreement in the nominalized relatives. In the remainder of the paper, we take up Lowe's suggestion and investigate how an analysis in terms of complex categories would play out.

4.2 Introducing Complex Categories

Complex categories were developed within ParGram (Butt et al. 1999) and implemented as part of XLE[7] in order to allow for a parameterization of syntactic categories. This parameterization enables the activation of one family of rules vs. others. In the English ParGram grammar complex categories are used to steer auxiliary selection (the "affix hopping" effects).[8] In the German grammar, complex categories are applied towards modeling parameters of how the verbal complex is realized. German is generally described as a V2 language, by which finite verbs in matrix clauses must appear in (roughly) second position and non-finite verbs (as well as verb particles) in clause final position. In embedded and relative clauses, on the other hand, all parts of the verbal complex are collected in the verb final position. The precise realization of the verbal complex differs according to the type of modals/auxiliaries contained within it and as to whether there is a coherent

[7] https://ling.sprachwiss.uni-konstanz.de/pages/xle/doc/notations.html#N3.4

[8] For an illustration, see the English grammar on the XLE-Web INESS site (https://clarino.uib.no/iness/xle-web) and try parsing *Helge had been having a nice day.*

verb such as *lassen* 'let', which disallows the *zu* 'to' infinitive. These lexical properties of verbs and auxiliaries/modals necessitate specialized rules for parts of the verbal complex, but scrambling possibilities of arguments and adjuncts or the overall licensing of arguments remain the same.

After much unsatisfactory experimentation with standard phrase structure rules to model the intricate details of German clause structure, the application of complex categories provided a computationally efficient and conceptually elegant way forward. In the current implementation verbs have a single entry for the stem. This stem specifies the subcategorization frame, case marking, compatibility with verbal particles and whether the item in question is a an auxiliary/modal [aux], a standard verb [v], or a verb with coherent properties [coh]. The inflectional morphology (coming out of a finite-state morphological analyzer; Schiller 1994) triggers a further parameterization according to: finite [fin], infinite [inf], participle [part].

In the syntax these lexical and morphological properties play out by allowing for rule parameterization through the formal tool of complex categories. Essentially, categories are "decorated" with a feature specification in square brackets, e.g., V[fin], V[inf], V[part]. One can add a feature declaration specifying legal values for a feature. Once the features are instantiated, they are not optional, that is, a feature cannot be left unspecified.

The current German ParGram grammar (Butt et al. 1999, Dipper 2003, Rohrer 2009)[9] assumes that verbs have two features: (_type, _infl) with `v, coh` and `aux` instantiating type, and `fin, inf` and `part` as values for infl.[10]

As determined by the lexicon and the morphology, a coherent finite verb, for example, is `V[coh,fin]`, while the participle of a standard verb is `V[v,part]`. This bottom up specification interacts with complex category rules in the syntax, triggering the appropriate syntactic behavior.

Let us begin with the matrix clause. The Cbar rule encompasses material from the finite verb onwards. This includes embedded complements. As the simplified version of the rule below shows, there are multiple possibilities. One is for a finite verb to be followed by a VP containing its arguments, the other is for a coherent verb to embed an XCOMP VP, the third accounts for the periphrastic *will* future, which requires a non-finite VP that could be headed by either a coherent or a standard verb, as seen in the `VPinf` rule.

```
Cbar --> { V[v,fin]   "either finite verb in single clause"
           VP
         | V[coh,fin] "or finite verb with XCOMP"
           VP: (^XCOMP = !)
         | V[aux,fin]  "or will-future"
           VPinf  }.
```

[9] See XLE-Web INESS website https://clarino.uib.no/iness/xle-web.

[10] The feature declaration is: `V[_type $ {v coh aux}, _infl $ {fin inf part}]`.

```
VPinf = { VP[v,inf]
        | VP[coh,inf] }.
```

The features specifying more details about the basic syntactic categories can function as variables which are instantiated as part of parsing. The VP rule is thus on the one hand very general, but on the other hand is prepared for features to be passed in from an outside rule activation, or for features to be instantiated by a lexical entry. For example, the third option in the Cbar rule could instantiate the (simplified) VP rule below as `VP[coh,inf]` as one possibility. In this case, the VP will call up the coherent version, as determined by checking for the feature `_type = coh`. This difference determines XCOMP embedding and also allows for recursive calls of VP embeddings.

```
VP[_type $ {v coh}, _infl $ {fin inf part}]   -->
         {  e: _type = v;
            @(VPconst ^)
         | e: _type = coh;
            @(VPconst (^XCOMP)) }
         { VC[_type,_infl]  "generic verbal complex"
         | VCflip[coh,fin]  "allow for auxiliary flip"}.
```

The final part of the rule above allows for either a generic verbal complex or for a special version with a flipped position of the auxiliary in embedded clauses. This is only possible with certain verbs, e.g., with coherent ones.

The introduction of the new formal tool of complex categories allowed for a new analytical perspective on well known intricate phenomena such as English auxiliary selection and German clause structure. Within the ParGram context, it was found that the introduction of complex categories allowed for conceptually cleaner analyses that were pleasingly coupled with computational efficiency as using complex categories is more efficient than performing f-structure checking on morphosyntactically determined features. In the following section, we turn to applying the concept of complex categories to an entirely different domain, namely mixed categories as found in Tamil nominalized complements and suggest that here too, complex categories open up a fruitful new analytical perspective.

5 Mixed Categories as Complex Categories

The intuition put forward in this section is to apply the possible parameterization of rule space to the problem of mixed categories by accumulating the features due to both derivational morphology and on-going historical change onto the major category. For example, we can model a gerund by assuming

that the main category is a V, but that it also carries a feature `n`, resulting in the mixed category V[n]. This models a composite category in which the V allows for the internal verbal licensing of arguments (nominative subject, etc.), but the [n] feature permits the simultaneous playing out of nominal features, such as case marking, perhaps projecting to an NP and, as a consequence, showing the external distribution of an NP. However, since the V[n] is not a full N, it can be limited to expressing a subset of nominal properties (e.g., no plural marking).[11]

We see the features on the complex categories as resulting from: 1) the effects of synchronic derivational morphology; 2) the effects of on-going diachronic reanalysis. As is well-known and discussed by Nikitina (2008) with respect to several case studies including verbal nouns and deverbal adjectives, one reason for the existence of mixed categories is gradual historical change by which lexical items are recategorized via reanalysis as they gradually accumulate more of the properties associated with one category rather than another. We analyze the Tamil complement patterns as classic cases of change in progress and posit complex C and V categories. We propose that complex categories provide a potentially elegant way of modeling gradual diachronic reanalysis by allowing for the definition of a possible parameter space which is affected by historical change and a coding of these parameters via features on complex categories, with attendant effects on the grammar.

5.1 Analysis of Nominalized Complements

The analysis we propose for (2), repeated below in (15), is shown in Figure 2. Our implementation was done within XLE by means of a small grammar of Tamil, which does not include a separate morphological analyzer and also does not do justice to Tamil's beautiful and complex orthography.[12] As such, we show the sublexical analysis simply as part of the c-structure and render the Tamil in a transliterated form.

We analyze the complement as being a CP which is headed by a C. This C is derived with the help of the original relativizer *-a* from an original quotative use of the verb 'say'. We do not provide a relative clause analysis of this at the featural level, but treat the *enp+a* as a combined form. This C has accumulated some nominal properties due to the incorporation of the pronoun, licensing the accusative case marking and triggering the external distribution of an NP, but not allowing for pluralization. The [nom] feature

[11] Our proposal bears similarities to Malouf's (2000) HPSG analysis of mixed categories in terms of inheritance hierarchies, by which a verbal noun, for example, can inherit both verbal and nominal properties. We are allowing the accumulation of mixed properties, but without invoking the formal restrictions and properties of inheritance hierarchies within the lexicon, see also Ash Asudeh and Toivonen (2008) for some discussion.

[12] We have implemented these as part of the larger Tamil ParGram grammar (Sarveswaran et al. 2018, 2019), which is also available on the INESS website.

on the complex category C percolates up to the CP because the instantiation of [nom] through the incorporated pronoun triggers the family of [nom] rules.

(15) [avan pizhai sey-tt-aan **enp-a-athu-ai**]
[PRON.3SM.NOM mistake do-PAST-3SM COMP-REL-PRON.3SN-ACC]
ram nirupi-tt-aan
Ram.NOM prove-PAST-3SM
'Ram proved (the fact) that he made a mistake.'

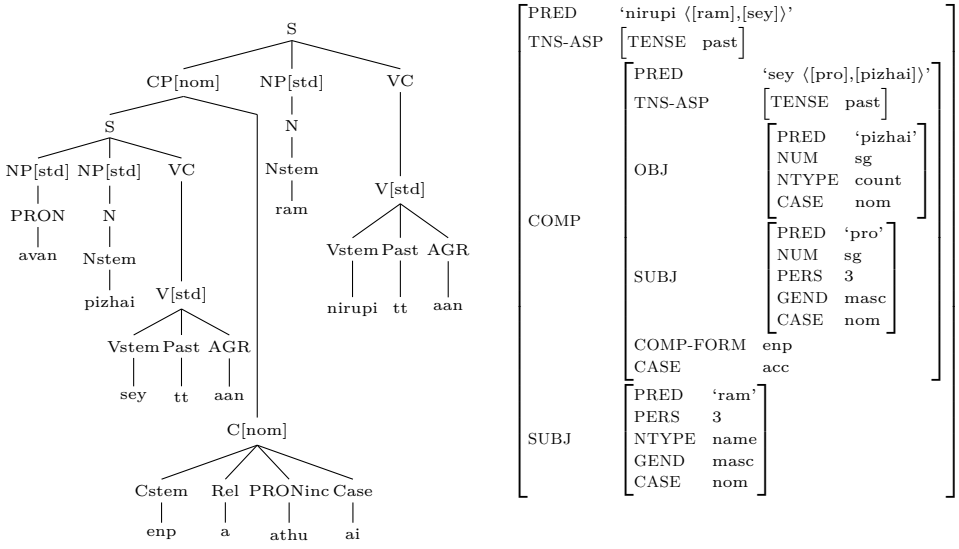


Figure 2: Complex Category Analysis of Nominalized Complement

The CP standardly contains an S, which is the default S found in the language and which exhibits all the scrambling properties (major constituents can scramble) of an S. The entire CP functions as a COMP, rendering a standard finite complementizer analysis at f-structure. The mixed category "oddities" of (16) in this case only play out in terms of the c-structure.[13]

5.2 Analysis of Nominalized Relative

The analysis of the nominalized relative (1), repeated below in (16), is along similar lines. We also posit a CP, but this CP has a c-structure that is analogous to that of a relative clause. The CP is headed by a V, as it would be in a RC. This V has been relativized, with the feature [rel] licensing the projection to the CP. The V has also been nominalized due to the incorporation of the pronoun, with this part of the feature licensing the accusative case marking and the external distribution as an NP, but prohibiting num-

[13]The NP and V carry the feature [std] (standard) vs. nominal [nom], verbal [v] or relative [rel]. Recall that once a type has been specified, it must always be instantiated.

ber marking. The nominalization is percolated up to the CP because the [nom] family of grammar rules is triggered by the nominal feature on the V. The relativization of the V means that subject-verb agreement cannot take place. But because the main category continues to be a V, all of the arguments predicated by the embedded verb can be realized with verbally licensed case. With respect to the f-structure, the embedded constituent functions as a COMP and is more in line with the f-structure analysis in Figure 2 than that of the f-structure analysis of relative clause in Figure 4.

(16) [avan pizhai **sey-t-a-athu-ai**]
[PRON.3SM.NOM mistake do-PAST-REL-PRON.3SN-ACC]
ram nirupi-tt-aan
Ram.3SM.NOM prove-PAST-3SM
'Ram proved (it) that he made a mistake.'

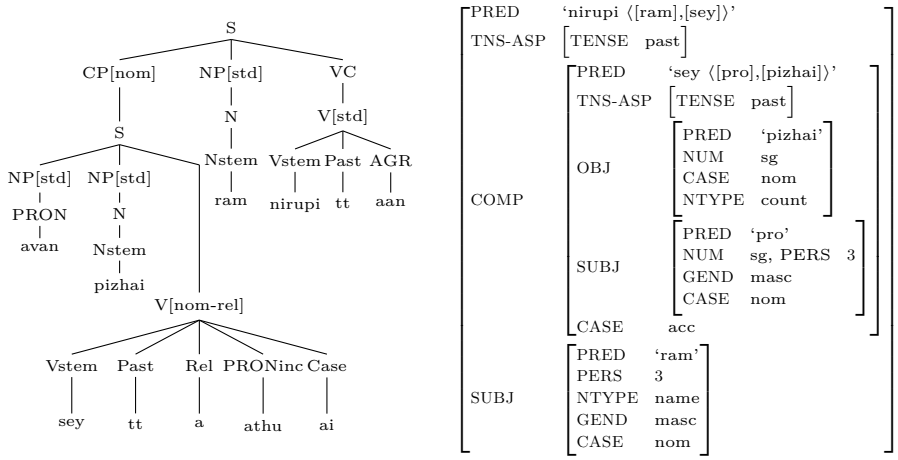


Figure 3: Complex Category Analysis of Nominalized Relative

For the sake of completeness, we also provide an analysis of the relative clause in (9-a), repeated below in (17). The relative clause modifies a head noun and is headed by a relativized verb. The [rel] feature is instantiated on the verb via the relativizer *-a* and percolates up to the CP because the [rel] on the V triggers the family of [rel] rules in the grammar via the complex category analysis.

(17) [angu nin-ṛ-a] paiyan-ai naan paar-t-en
there stand-PAST-REL boy-ACC I.NOM.1S see-PAST-1S
'I saw the boy who stood there.'

The f-structure analysis follows the standard ParGram analysis of relative clauses so that it is represented as an adjunct modifying the head noun 'boy', with a 'pro' functioning as the subject of the relative clause.

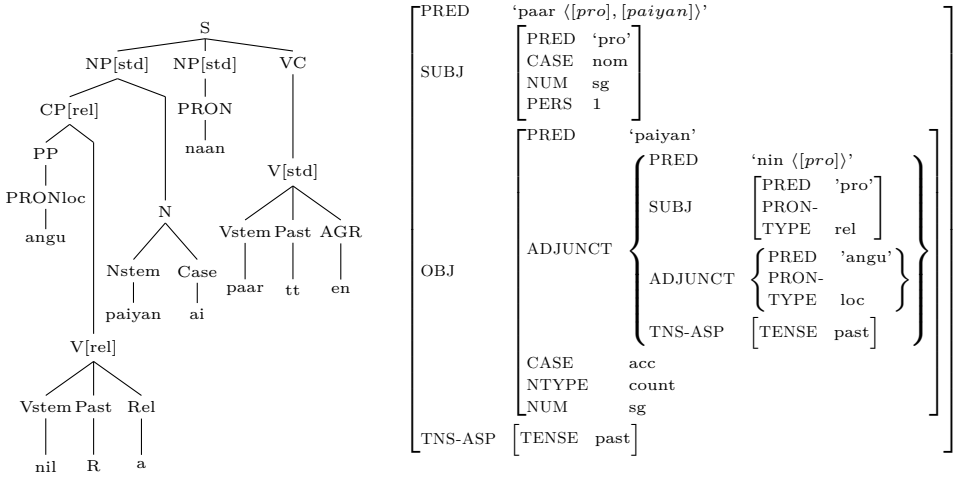


Figure 4: C-structure and F-structure for a Relative Clause

6 Conclusion

This paper has presented an analysis of Tamil nominalized complements. We have identified them as a type of mixed category, whereby the nominalization is due to the incorporation of a pronoun into the head of the CP. We analyzed two different constructions, one containing a complementizer that is related to a quotative use of the verb 'say'. Both constructions feature relativization, which seems to be a basic way forming embedded nominal complement clauses in Tamil.

We proposed to analyze the complicated morphology found on the (originally) verbal stems in terms of complex categories, with the intuition being that the mixed properties of syntactic categories can be modeled through features on a syntactic category such as V or C. This allows for a parameterization of the grammar rules according to these features and also allows a projection of a CP from a relativized V or the projection of an NP from a nominalized V.

References

Amritavalli, R. 2013. Nominal and interrogative complements in Kannada. In Howard Lasnik, Myung-Kwan Park, Yoichi Miyamoto, Daiko Takahashi, Hideki Maki, Masao Ochi, Koji Sugisaki and Asako Uchiboro (eds.), *Deep insights, broad perspectives: Essays in honor of Mamoru Saito*, pages 1–21, Kaitakusha.

Amritavalli, R. and Jayaseelan, K.A. 2008. Finiteness and Negation in Dravidian. In Guglielmo Cinque and Richard S. Kayne (eds.), *The Oxford Handbook of Comparative Syntax*, Oxford: Oxford University Press.

Arden, Albert Henry. 1962. *A progressive grammar of common Tamil*. Christian Literature Society.

Ash Asudeh, Mary Dalrymple and Toivonen, Ida. 2008. Constructions with Lexical Integrity. *Journal of Language Modeling* 1(1), http://dx.doi.org/10.15398/jlm.v1i1.56.

Balusu, Rahul. 2020. The Quotative Complementizer Says "I'm too Baroque for that". In *Formal Approaches to South Asian Languages*, volume 3.

Bögel, Tina. 2015. *The Syntax–Prosody Interface in Lexical Functional Grammar*. Ph. D.thesis, University of Konstanz, Konstanz.

Bresnan, Joan. 1976. On the Form and Functioning of Transformations. *Linguistic Inquiry* 7(1), 3–40.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Bresnan, Joan, Asudeh, Ash, Toivonen, Ida and Wechsler, Stephen. 2016. *Lexical Functional Syntax*. Oxford: Wiley-Blackwell, second edition. First edition by Joan Bresnan, 2001, Blackwell.

Bresnan, Joan and Mugane, John. 2006. Agentive nominalizations in Gĩkũyũ and the theory of mixed categories. In *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, pages 201–234, Stanford, CA: CSLI Publications.

Butt, Miriam, King, Tracy Holloway, Niño, María Eugenia and Segond, Frédérique. 1999. *A Grammar Writer's Cookbook*. Stanford, CA: CSLI Publications.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Corbett, Greville. 2006. *Agreement*. Cambridge: Cambridge University Press.

Corbett, Greville. 2007. Canonical typology, suppletion and possible words. *Language* 83, 8–42.

Croft, William A. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: Chicago University Press.

Crouch, Richard, Dalrymple, Mary, Kaplan, Ronald M., King, Tracy Holloway, Maxwell III, John T. and Newman, Paula. 2017. XLE Documentation. Technical Report, Palo Alto Research Center (PARC), https://ling.sprachwiss.uni-konstanz.de/pages/xle/doc/xle_toc.html.

Dipper, Stefanie. 2003. *Implementing and Documenting Large-Scale Grammars — German LFG*. Ph. D.thesis, IMS, University of Stuttgart.

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*, pages 29–32, Association for Computational Linguistics.

Jurafsky, Daniel and Martin, James. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguis-*

tics, and Speech Recognition, Second Edition. Upper Saddle River, NJ: Prentice Hall.

Klamer, Marian. 2000. How report verbs become quote markers and complementisers. *Lingua* 110, 69–98.

Krishnamurti, Bhadriraju. 2003. *The Dravidian Languages*. Cambridge University Press.

Lahiri, Aditi, Jongman, Allard and Sereno, Joan. 1990. The pronomincal clitic [dər] in Dutch: a theoretical and experimental approach. *Yearbook of Morphology* 3, 115–127.

Lehmann, Thomas. 1993. *A grammar of modern Tamil*. Pondicherry Institute of Linguistics and Culture.

Lowe, John. 2016. Participles, gerunds and syntactic categories. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King and Stefan Müller (eds.), *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, pages 401–421, Stanford, CA: CSLI Publications.

Malouf, Robert. 2000. *Mixed Categories in the Hierarchical Lexicon*. Stanford, CA: CSLI Publications.

Nikitina, Tatiana V. 2008. *The Mixing of Syntactic Properties and Language Change*. Ph. D.thesis, Stanford University.

Nikitina, Tatiana V. and Haug, Dag T.T. 2016. Syntactic Nominalization in Latin: a case of non-canonical subject agreement. *Transactions of the Philological Society* 114, 25–50.

Nikolaeva, Irina and Spencer, Andrew. 2020. *Mixed Categories: The Morphosyntax of Noun Modification*. Cambridge: Cambridge University Press.

Rajendran, S. 2001. Typology of nominalization in Tamil. *Language in India* 1(7).

Rajendran, S. 2012. Preliminaries To The Preparation Of A Spell And Grammar Checker For Tamil, https://www.academia.edu/12504639/PRELIMINARIES_TO_THE_PREPARATION_OF_A_SPELL_AND_GRAMMAR_CHECKER_FOR_TAMIL, accessed Nov. 3, 2020.

Rohrer, Christian. 2009. Problems of German VP Coordination. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG09 Conference*, Stanford,CA: CSLI Publications.

Rosén, Victoria, De Smedt, Koenraad, Meurer, Paul and Dyvik, Helge. 2012. An open infrastructure for advanced treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29.

Sarveswaran, K, Dias, Gihan and Butt, Miriam. 2018. ThamizhiFST: A Morphological Analyser and Generator for Tamil Verbs. In *2018 3rd International Conference on Information Technology Research (ICITR)*, pages 1–6, IEEE.

Sarveswaran, K, Dias, Gihan and Butt, Miriam. 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*, pages 76–86, Dresden, Germany: Association for Computational Linguistics.

Schiffman, Harold. 1969. *A Transformational grammar of the Tamil Aspectual system*. University of Chicago: University of Washington.

Schiller, Anne. 1994. DMOR - User's Guide. Technical Report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Selkirk, Elisabeth O. 1995. The prosodic structure of function words. In Jill N. Beckmann, Laura W. Dickey and Suzanne Urbanczyk (eds.), *Papers in Optimality Theory*, University of Massachusetts: Department of Linguistics.

Spencer, Andrew. 2015. Participial relatives in LFG. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG15 Conference*, pages 378—398, Stanford, CA: CSLI Publications.

Stowell, Tim. 1981. *Origins of Phrase Structure*. Ph. D.thesis, MIT.

Sulger, Sebastian, Butt, Miriam, King, Tracy Holloway, Meurer, Paul, Laczkó, Tibor, Rákosi, György, Dione, Cheikh Bamba, Dyvik, Helge, Rosén, Victoria and De Smedt, Koenraad. 2013. ParGramBank: The ParGram parallel treebank. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 550–560.

Zvelebil, Kamil. 1974. *Tamil Literature*. Otto Harrassowitz Verlag.