INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

TAMIL TREEBANK GUIDELINES - V 0.2

Project Name: **A Syntactic Parser for Tamil: A Data-driven Approach**

Funding Agency: **Tamil Virtual Academy (TVA)**

Letter No :Lr.N0.TVA/TC-EOI /2022/006-5, Dated. 05.01.2023

Report Submitted by,

Dr. Parameswari Krishnamurthy (PI of the Project)
Language Technology Research Center(LTRC),
International Institute of Information Technology, Hyderabad
Prof. C R Rao Road,
Gachibowli, Hyderabad 500 032,

Telangana, INDIA, Phone: +91 85000 25207, email: param.krishna@iiit.ac.in

Tamil Universal Dependency Relations

	Nominals	**Clauses**	**Modifier Words**	**Function words**
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative	advcl	advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith	punct root

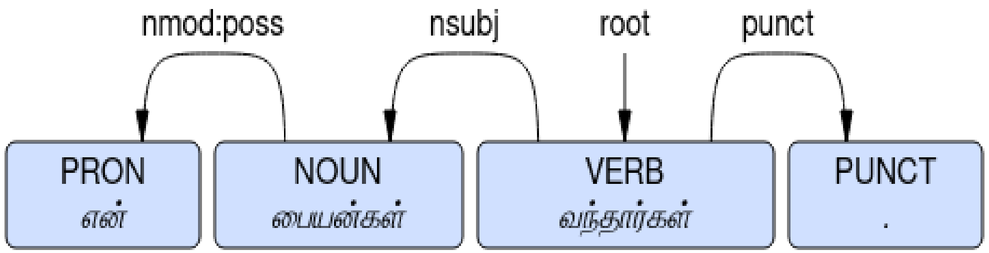
Core arguments

1. nsubj

A nominal subject (nsubj) is a nominal which is the syntactic subject and the proto-agent of a clause. That is, it is in the position that passes a typical grammatical test for subjecthood, and this argument is the more agentive, the do-er, or the proto-agent of the clause.

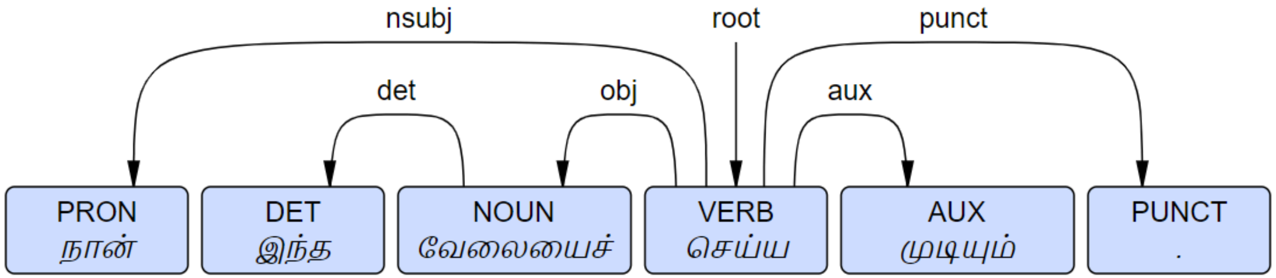
1.1 **nsubj in nominative case**

(1) என் பையன்கள் வந்தார்கள் .

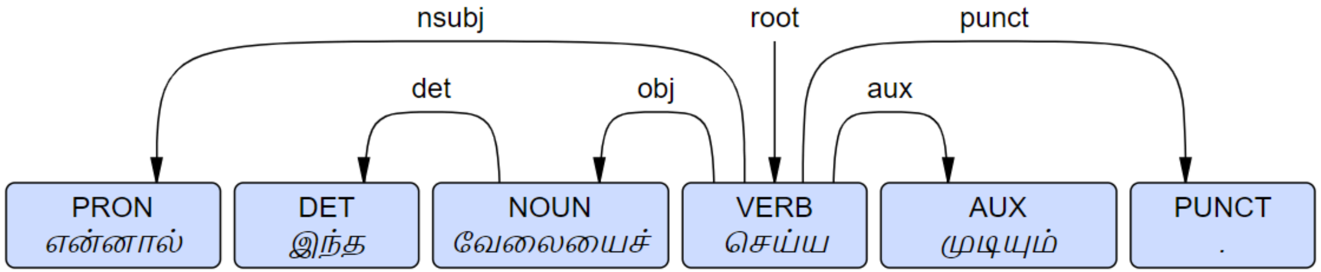


However, when the predicate indicates the capabilitative mood, the subject is optionally marked for the *Instrumental case* marker and the verb gets default agreement i.e. third person-neuter

(2) நான் இந்த வேலையைச் செய்ய முடியும் .

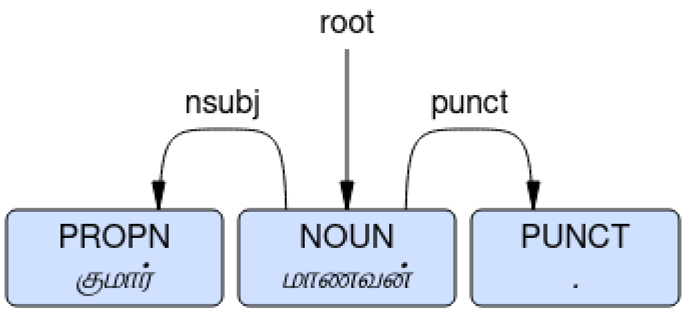


(3) என்னால் இந்த வேலையைச் செய்ய முடியும் .

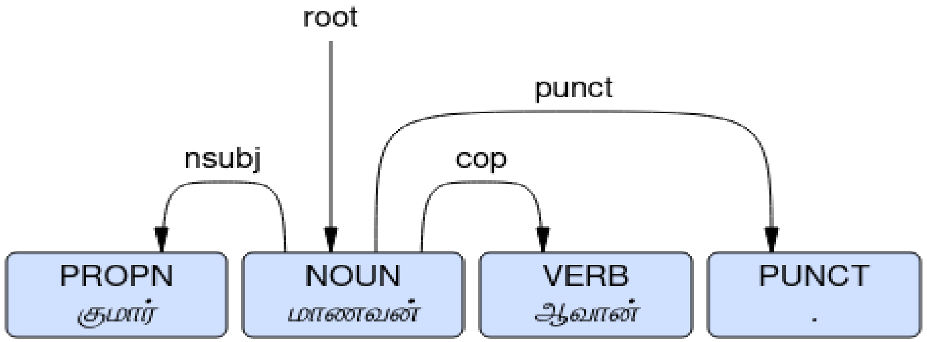


The head of **nsubj** is not always a verb in Tamil when the nominal and adjectival predicates occur optionally with the copula verb ‘ஆகு‘. In such cases non-verbal predicates are considered as head (i.e. root) and the copula verb if it is present is related as **aux** (auxiliary) to the **root.**

(4) குமார் மாணவன் .



(5) குமார் மாணவன் ஆவான்



Certain predicates require their nsubj to be case-marked by non-nominative case markers such as the dative, locative and instrumental markers (like in (3)) in Tamil.

1.2 ‘nsubj’ in Dative Subject Constructions:

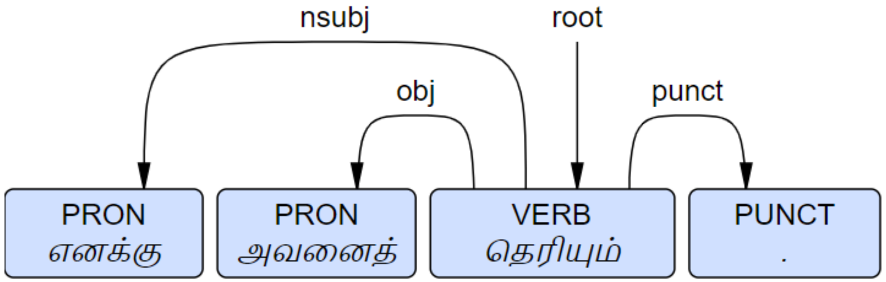
The dative marked subject acts as an *experiencer* subject and the verb agrees with the object. In Tamil, stative predicates expressing the notion of mental, emotional and physical experience require the case marking pattern of DAT-ACC (Lehmann, 1993:180).

1.2.1 Verbs of mental experience

When the verbs such as தெரி (teri) ‘know’ and புரி (puri) ‘understand’ in Tamil, the *logical* subject is marked with the dative case marker.

Note: there is a **default agreement** with the verb.

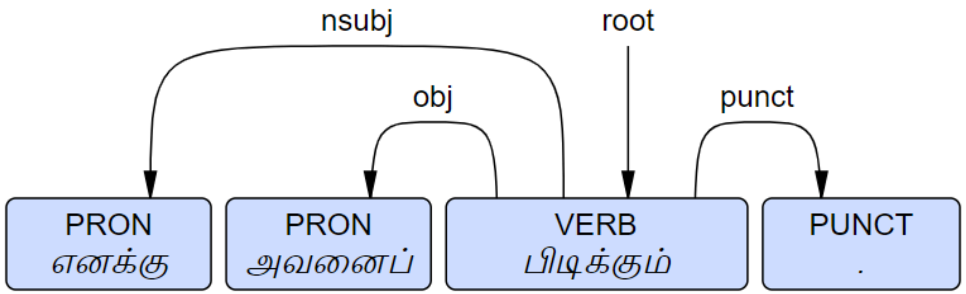
(6) எனக்கு அவனைத் தெரியும் .



1.2.2 Verbs of emotional experience

Verbs like பிடி (piti) (tr.) ‘like’ etc., in Tamil express emotional experience with the dative-marked subject.

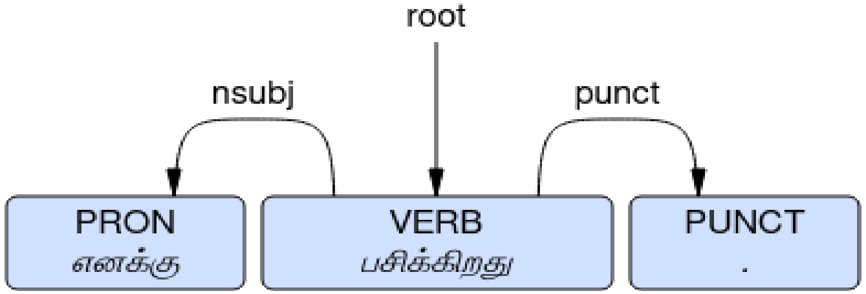
(7) எனக்கு அவனைப் பிடிக்கும் .



1.2.3 Verbs of psycho-semantic, physical and physiological experiences

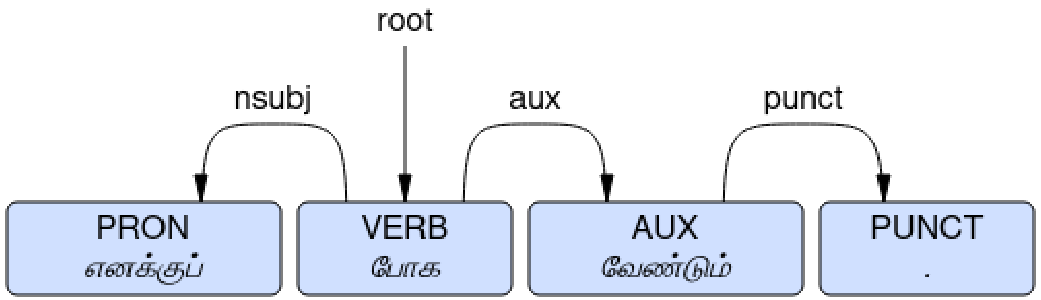
Verbs such as பசி (paci) ‘be hungry’, வலி (vali) ‘be painful’, and அரி (ari) ‘be itching’ in Tamil conveys psycho-semantic, physical and physiological experiences and requires their subject with the dative marker.

(8) எனக்கு பசிக்கிறது .



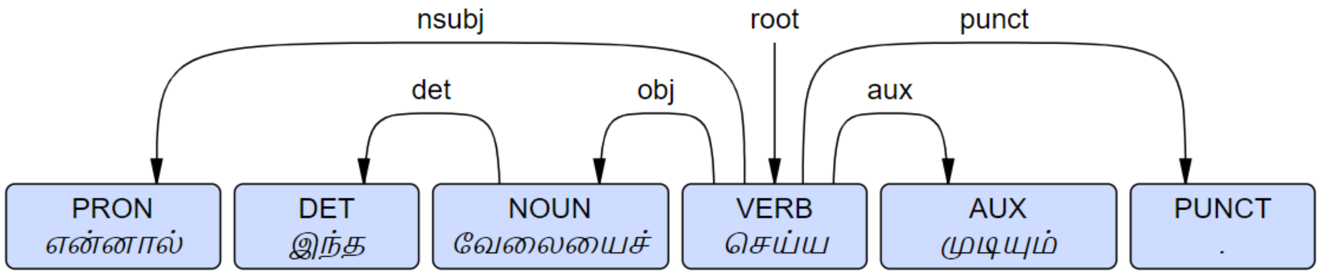
Auxiliaries like ‘வேண்டும்’ in Tamil require the subject in the dative case marker.

(9) எனக்குப் போக வேண்டும்.



1.3. ‘nsubj’ in Instrumental Subject Constructions

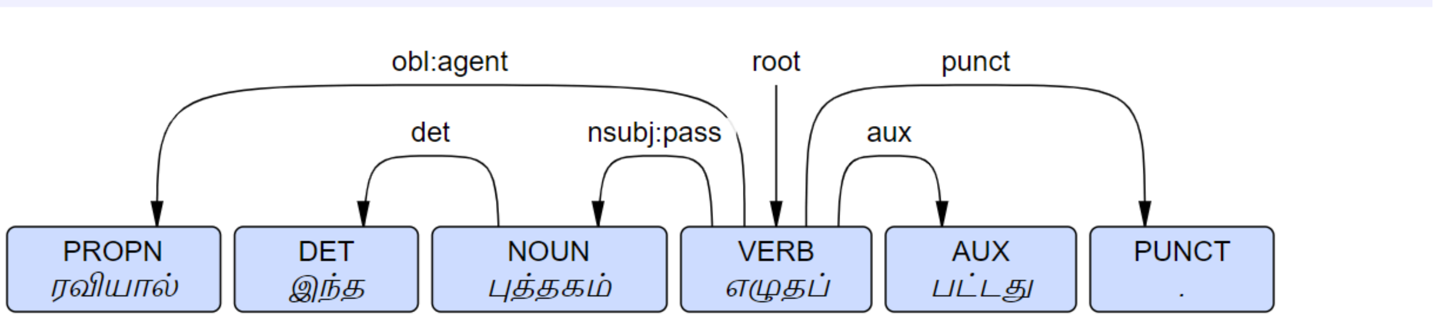
(10) என்னால் இந்த வேலையைச் செய்ய முடியும் .



1.4. ‘nsubj’ in Passive Construction (nsubj:pass)

A passive nominal subject is a noun phrase which is the syntactic subject of a passive clause.

(11) ரவியால் இந்த புத்தகம் எழுதப்பட்டது.



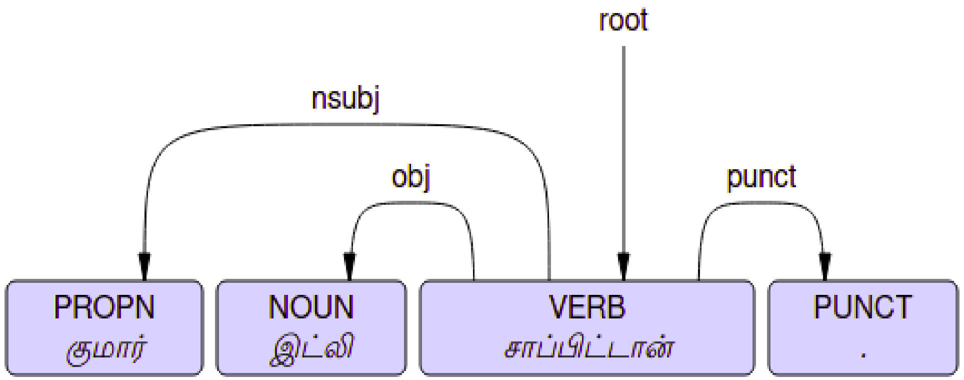
2. obj: object

The object of a verb is the second most core argument of a verb after the subject. Typically, it is the noun phrase that denotes the entity acted upon or which undergoes a change of state or motion (the proto-patient) (from UD annotation guidelines).

In Tamil, the rational objects will be marked by the accusative case explicitly, and irrational objects are optionally marked by the accusative case.

(12) ரவி கமலாவைப் பார்த்தான்

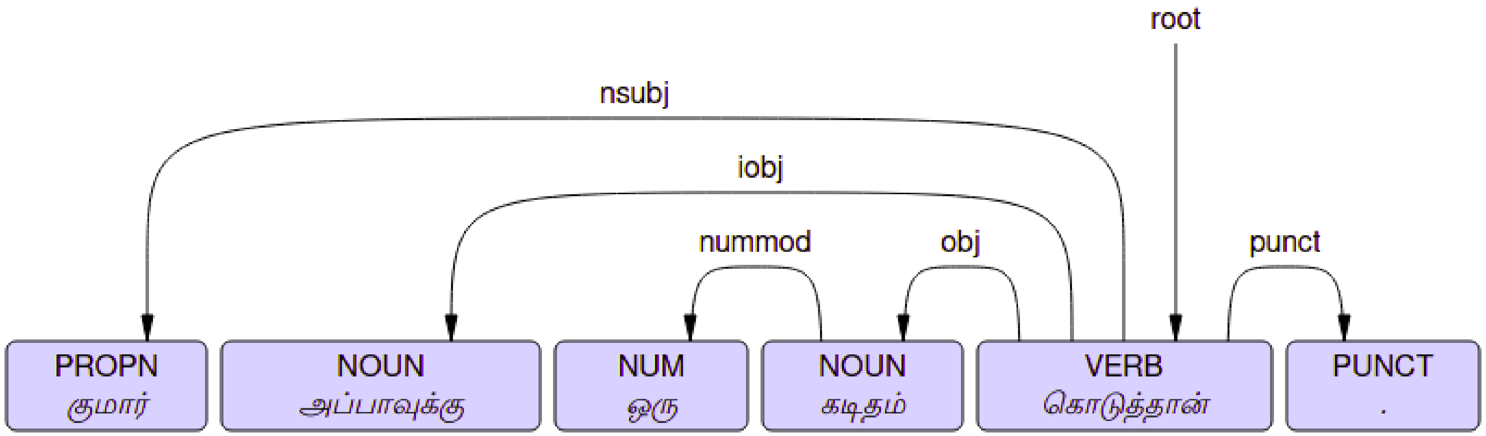
(13) குமார் இட்லி சாப்பிட்டான் .



3. iobj: indirect object

The noun phrase which is the recipient of a ditransitive verb i.e. indirect object is marked as ‘iobj’.

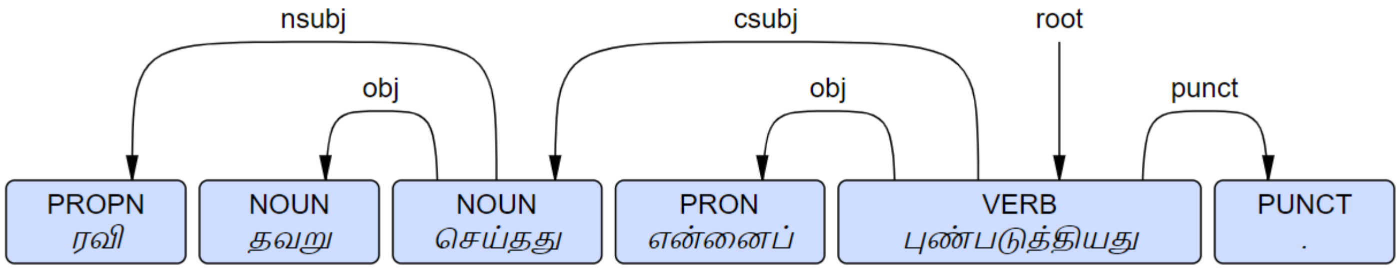
(14) குமார் அப்பாவுக்கு ஒரு கடிதம் கொடுத்தான் .



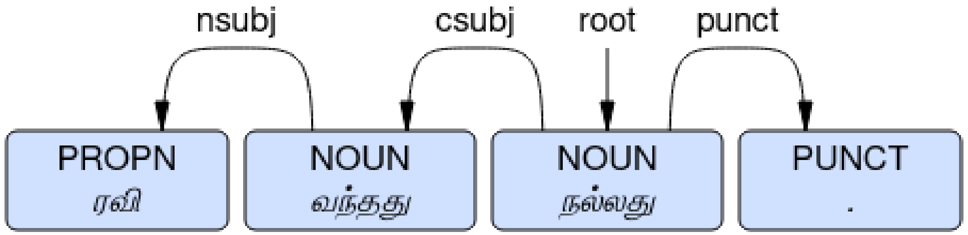
4. csubj: clausal subject

When the subject is realised as a clause, it is marked as ‘csubj’. The root of the sentence can either be a verb or a noun in case of copular construction.

(15) ரவி தவறு செய்தது என்னைப் புண்படுத்தியது



(16) குமார் வந்தது நல்லது.

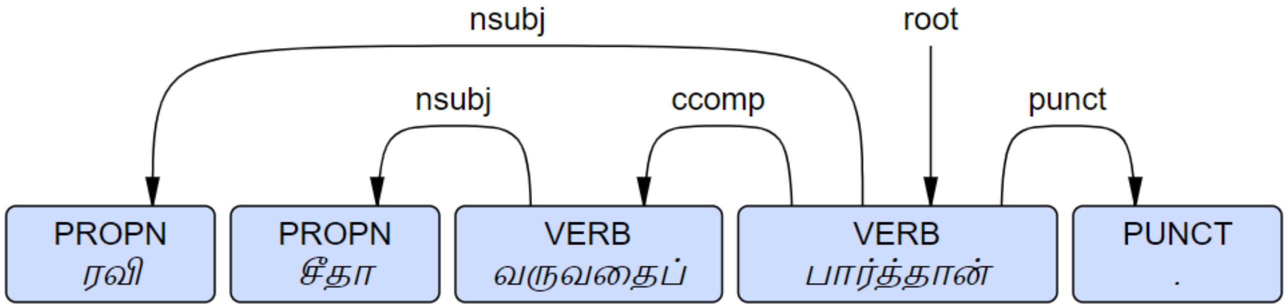


5. ccomp: clausal complement

When the object is realised as a clause, it is marked as 'ccomp'.

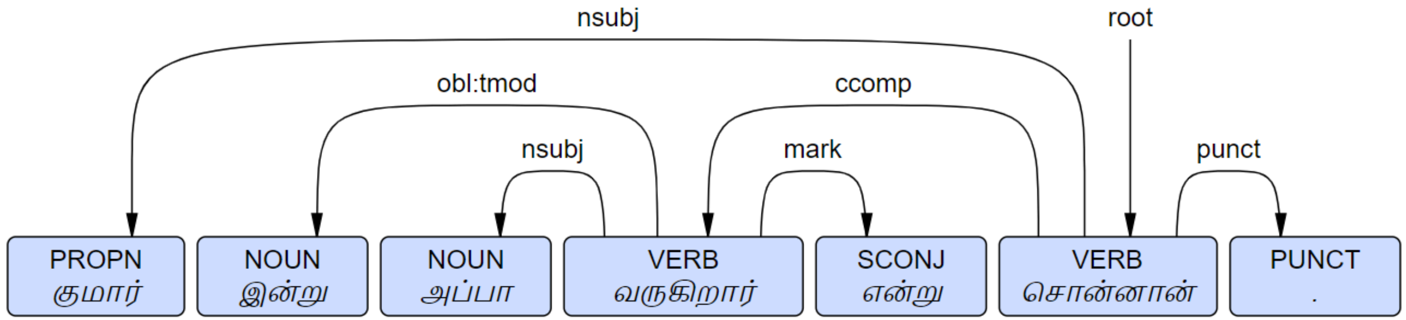
(17) ரவி சீதா வருவதைப் பார்த்தான்

This is also ccomp, and comp is marked by the வருவதை itself. For the below two examples, it is marked externally.



In certain cases, the complementizer (i.e mark) links the clausal object with the root.

(18) குமார் இன்று அப்பா வருகிறார் என்று சொன்னான்.

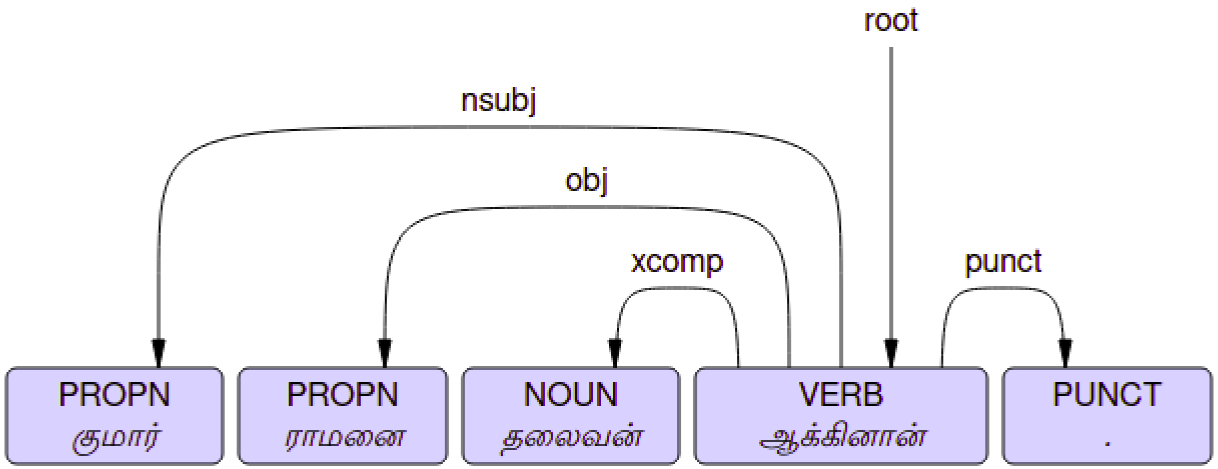


6. xcomp: open clausal complement

An open clausal complement without its own subject is marked with the tag xcomp. The reference of the subject to the xcomp is controlled by either subject or object of the next higher clause.

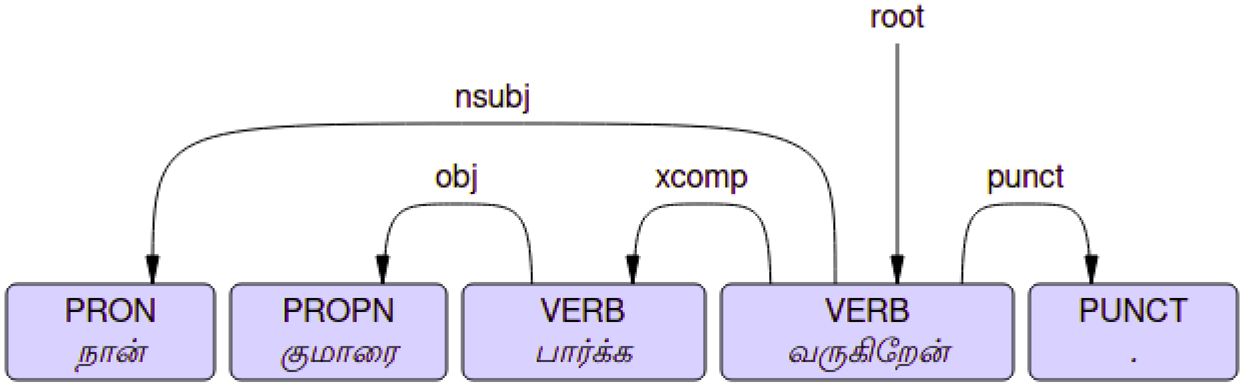
6.1. 'xcomp' can be a nominal complement.

(19) குமார் ராமனை தலைவன் ஆக்கினான்.



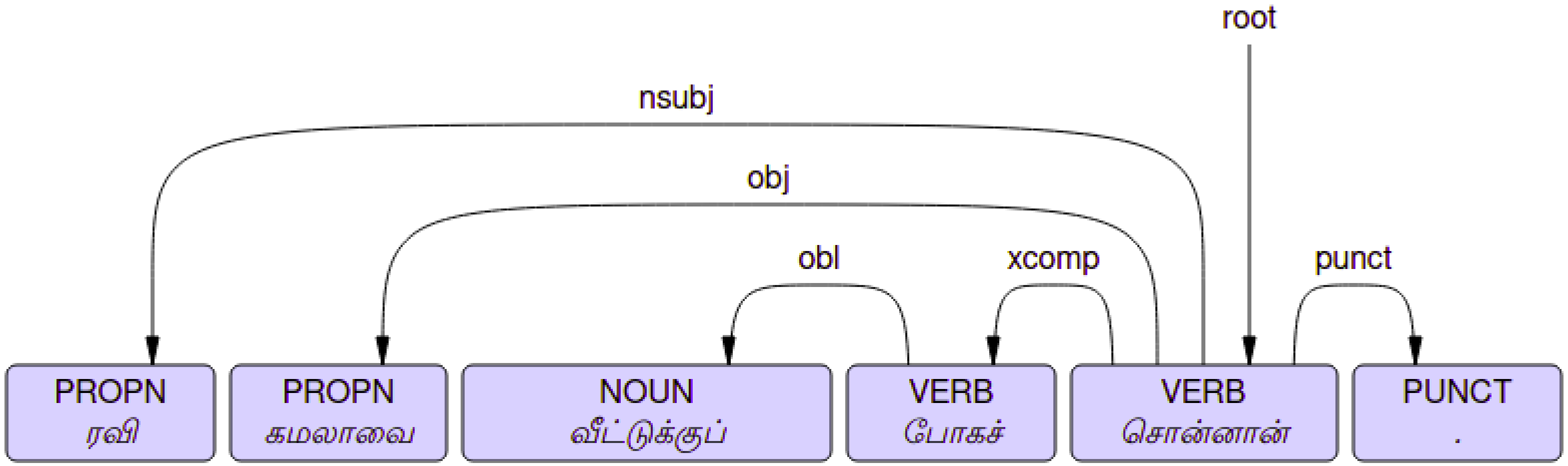
6.2. ‘xcomp’ can be a verbal complement

(i) Subject control
(20) நான் குமாரை பார்க்க வருகிறேன் .



(ii) Object control

(21) ரவி கமலாவை வீட்டுக்குப் போகச் சொன்னான்.

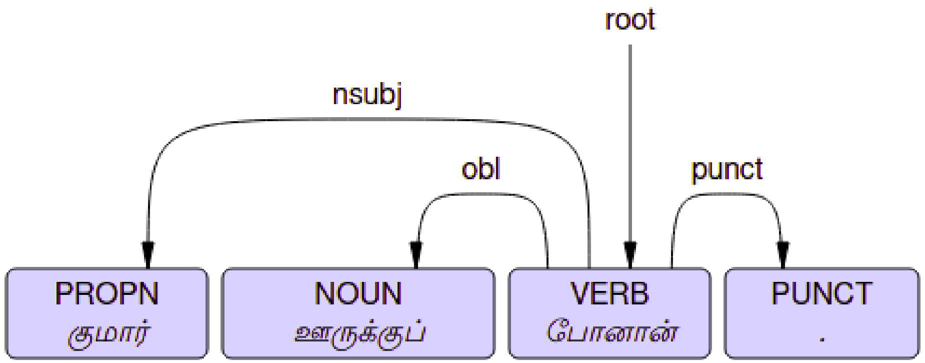


Non-core dependents

7. obl

The obl relation is used for a nominal (noun, pronoun, noun phrase) functioning as a non-core (oblique) argument or adjunct.

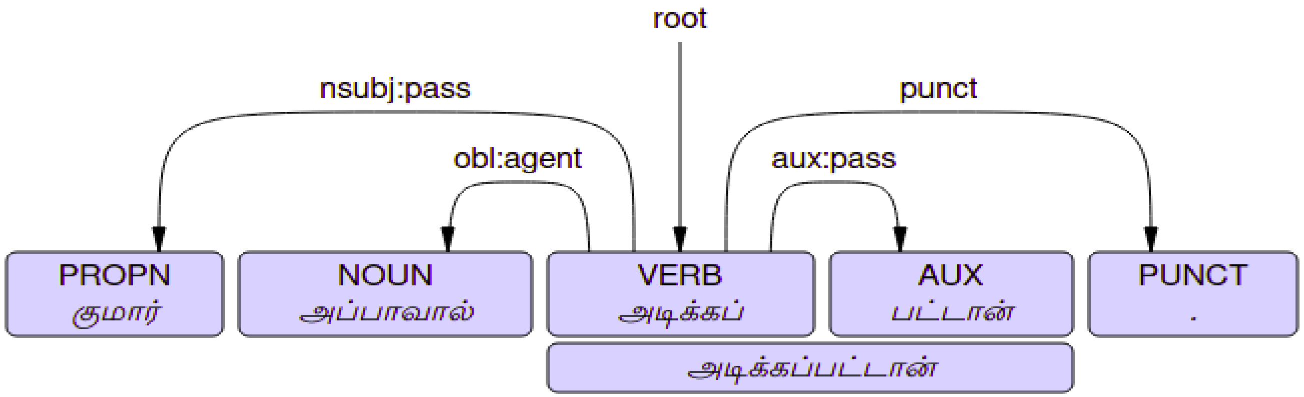
(22) குமார் ஊருக்குப் போனான் .



7.1. obl: agent

The relation obl:agent is used for nouns which are agents of passive verbs.

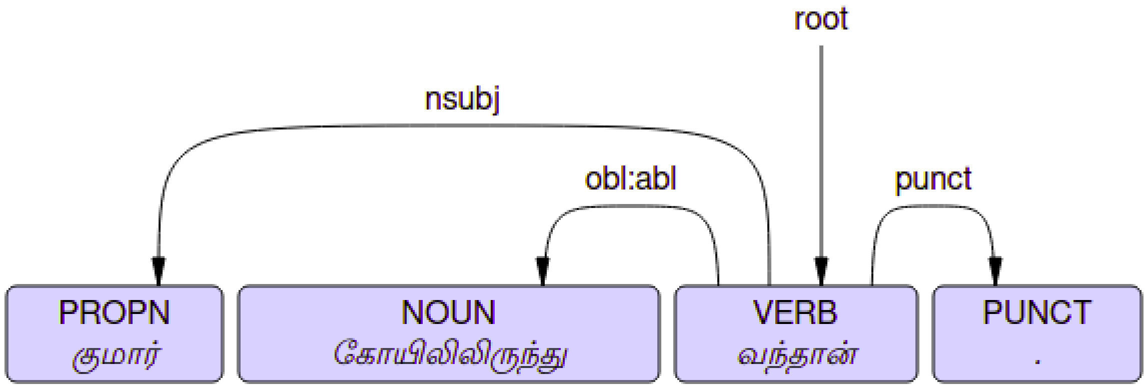
(23) குமார் அப்பாவால் அடிக்கப்பட்டான் .



7.3. obl:abl

The relation obl:abl is used for nouns which express the source of place

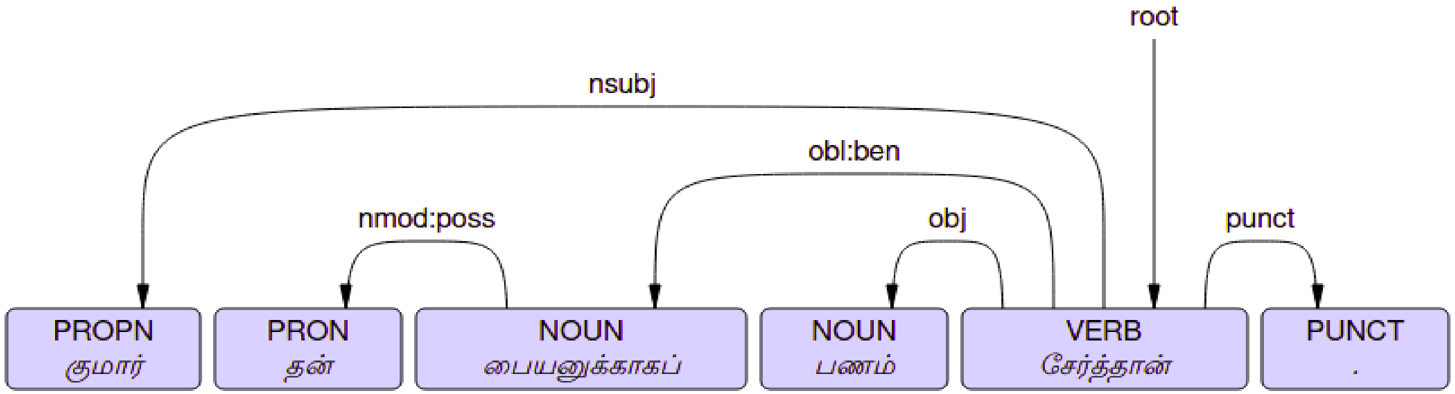
(24) குமார் கோயிலிலிருந்து வந்தான் .



7.4. obl:ben

The relation obl:ben is used for nouns which are benefactors of actions.

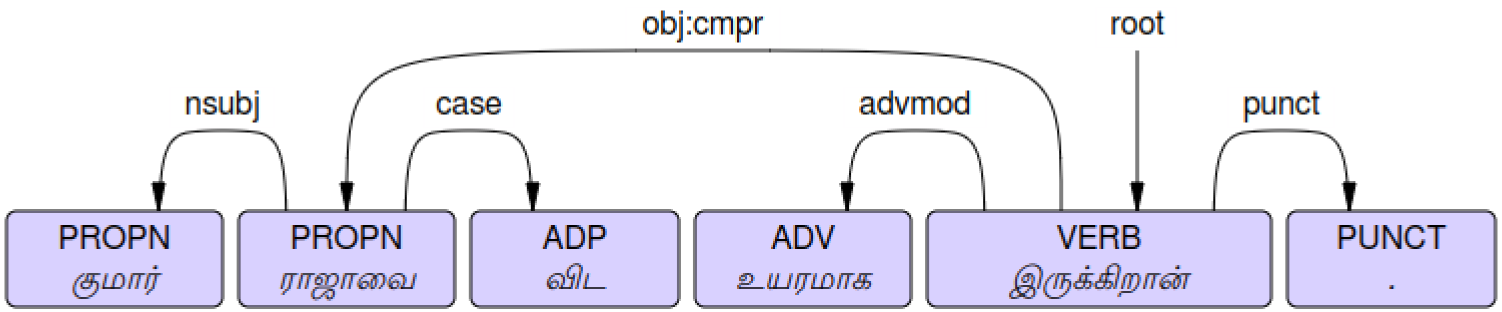
(25) குமார் தன் பையனுக்காகப் பணம் சேர்த்தான் .



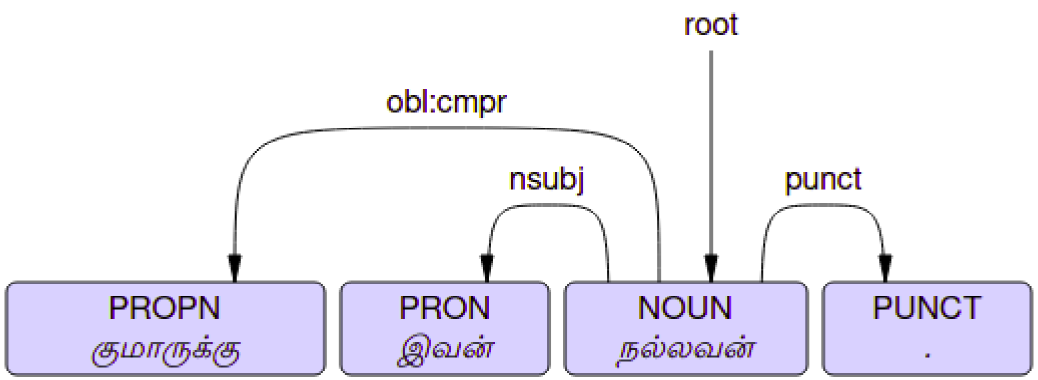
7.5. obl:cmpr

The relation obl:cmpr is used for nouns which are used as comparand.

(26) குமார் ராஜாவை விட உயரமாக இருக்கிறான் .



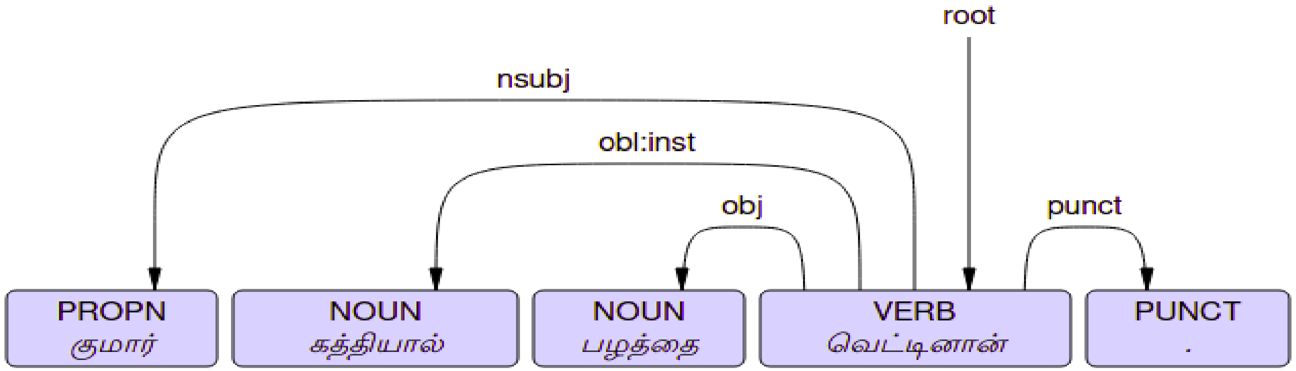
(27) குமாருக்கு இவன் நல்லவன் .



7.6. obl:inst

The relation ‘obl:inst’ is used for nouns which are instruments for actions.

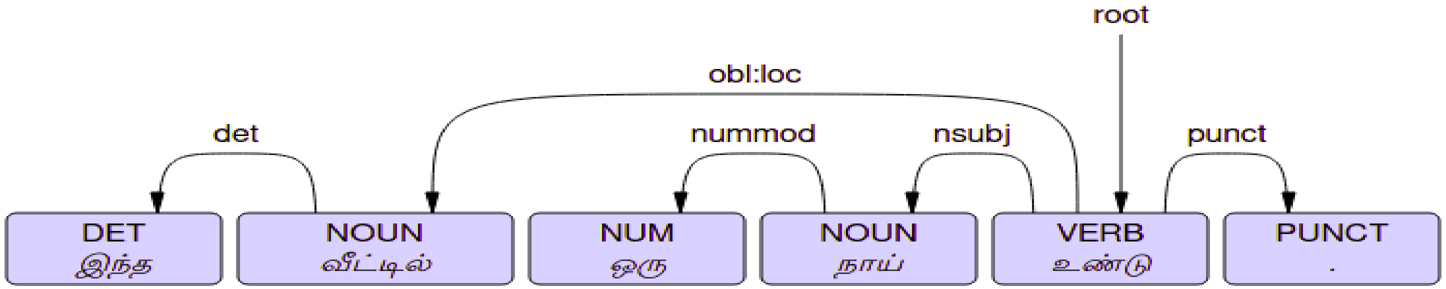
(28) குமார் கத்தியால் பழத்தை வெட்டினான் .



7.7. obl:loc

The relation ‘obl:loc’ is used for nouns which shows location (locative).

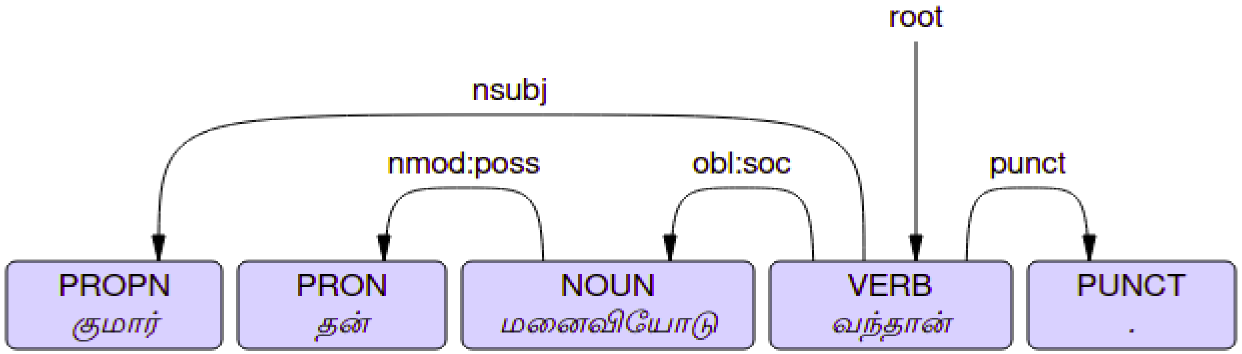
(29) இந்த வீட்டில் ஒரு நாய் உண்டு .



7.8. obl:soc

The relation ‘obl:soc’ is used for nouns which shows association (associative/sociative).

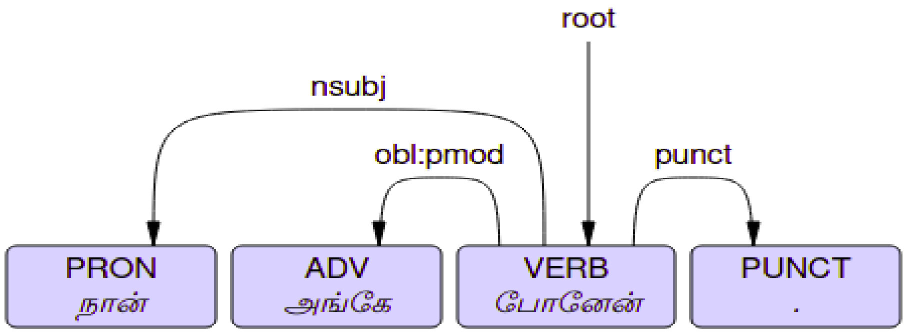
(30) குமார் தன் மனைவியோடு வந்தான் .



7.9. obl:pmod

A place modifier is a subtype of the ‘obl’ relation. It is tagged ‘pmod’ if the modifier is specifying a place.

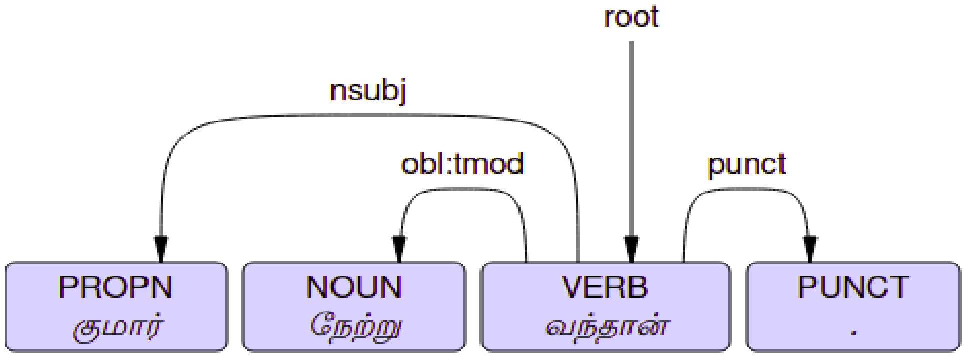
(31) நான் அங்கே போனேன் .



7.10. obl:tmod

A temporal modifier is a subtype of the ‘obl’ relation. If the modifier is specifying a time, then it is labelled as ‘tmod’.

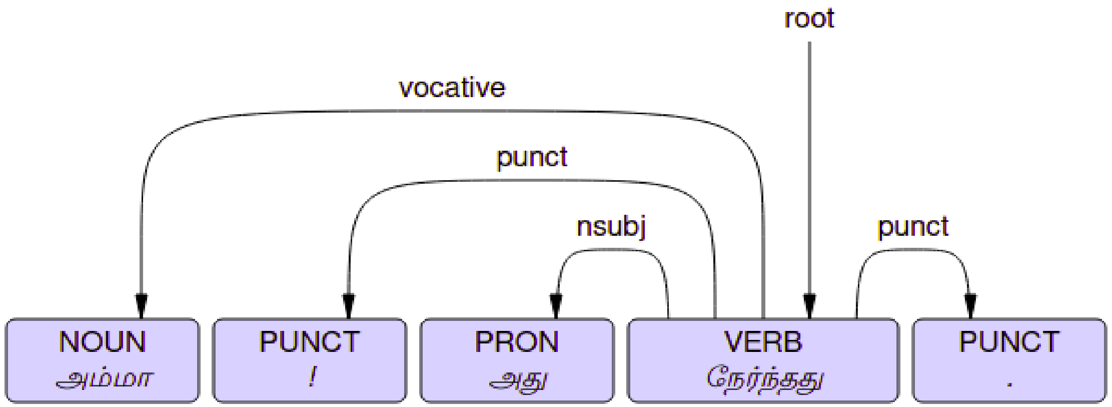
(32) குமார் நேற்று வந்தான் .



8. vocative: vocative

The *vocative* relation is used to mark a dialogue participant addressed in a text (common in conversations, dialogue, emails, newsgroup postings, etc.). The relation links the addressee’s name to its host sentence.

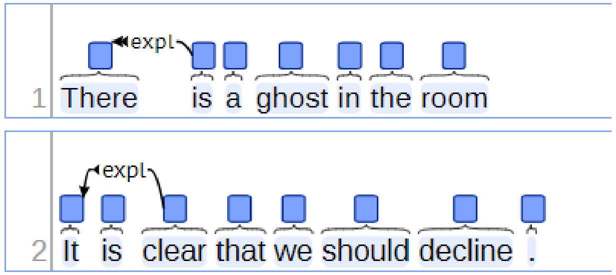
(33) அம்மா! அது நேர்ந்தது .



9. expl

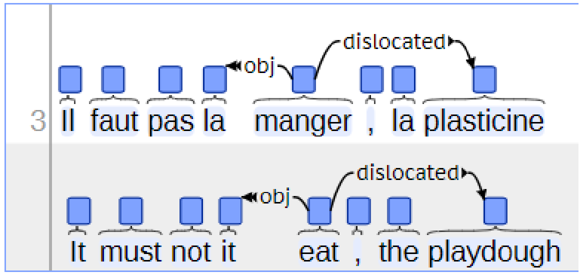
This relation captures expletive or pleonastic nominals. These are nominals that appear in an argument position of a predicate but which do not themselves satisfy any of the semantic roles of the predicate. The main predicate of the clause (the verb or predicate adjective or noun) is the governor.

Example: **There** is a ghost in the room.


expl
1 There is a ghost in the room
expl
2 It is clear that we should decline .

10. dislocated: dislocated elements

The ‘dislocated’ relation is used for fronted or postposed elements that do not fulfil the usual core grammatical relations of a sentence. These elements often appear to be in the periphery of the sentence, and may be separated off with a comma intonation.


dislocated
obj
3 Il faut pas la manger , la plasticine
dislocated
obj
It must not it eat , the playdough

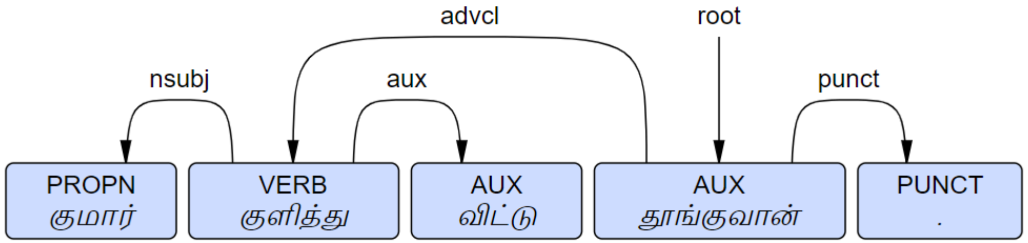
11. advcl

An adverbial clause modifier (advcl) is a clause which modifies a verb or other predicate (adjective, etc.), as a modifier not as a core complement.

This includes things such as a temporal clause, consequence, conditional clause, purpose clause, etc.

The dependent must be clausal (or else it is an advmod) and the dependent is the main predicate of the clause.

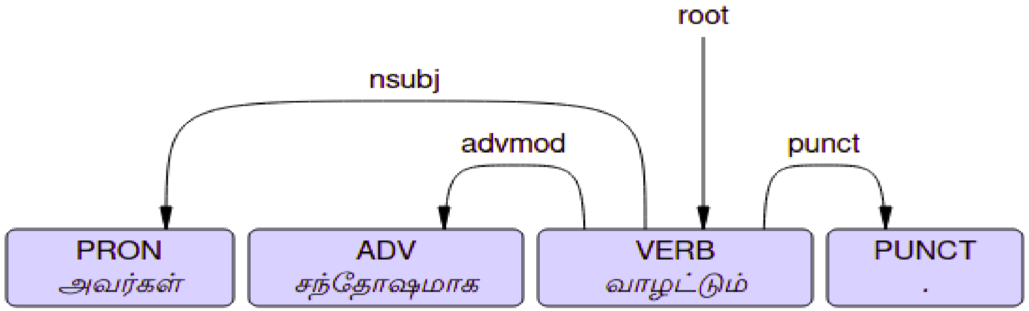
(34) குமார் குளித்து விட்டு தூங்குவான் .



12. advmod

An adverbial modifier (advmod) of a word is a (non-clausal) adverb or adverbial phrase that serves to modify a predicate or a modifier word.

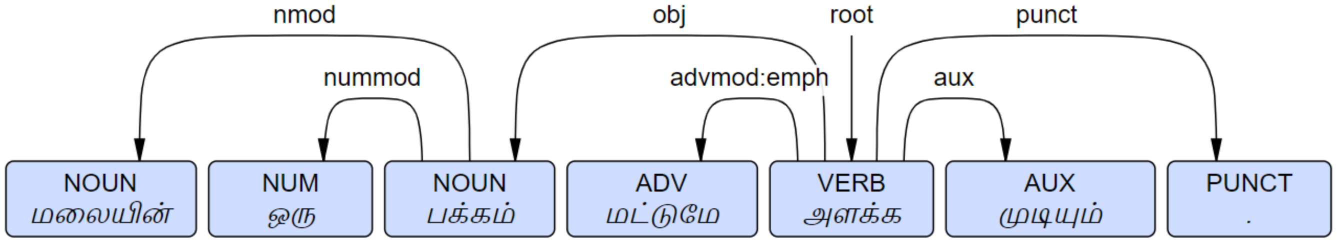
(35) அவர்கள் சந்தோஷமாக வாழட்டும் .



12.1. advmod:emph

A limited set of adverbs can also modify nominals (e.g., ***only*** *on Monday*). The advmod relation or its subtype has to be used in such cases, too

(36) மலையின் ஒரு பக்கம் மட்டுமே அளக்க முடியும் .



13. discourse: discourse element

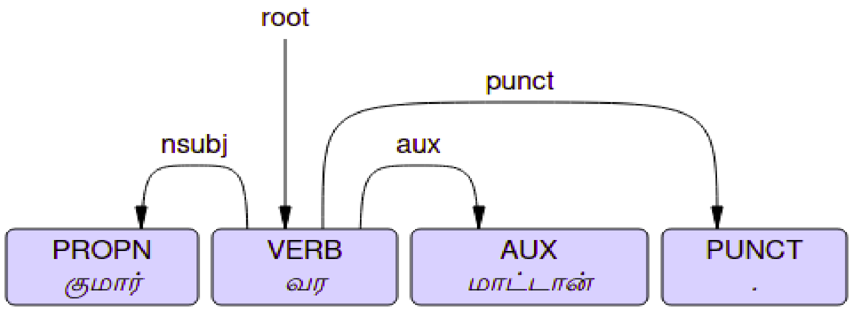
This is used for interjections and other discourse particles and elements (which are not clearly linked to the structure of the sentence, except in an expressive way).

Example: He is smiling **:)**

14. aux

An aux (auxiliary) of a clause is a function word associated with a verbal predicate that expresses categories such as tense, mood, aspect, voice or evidentiality.

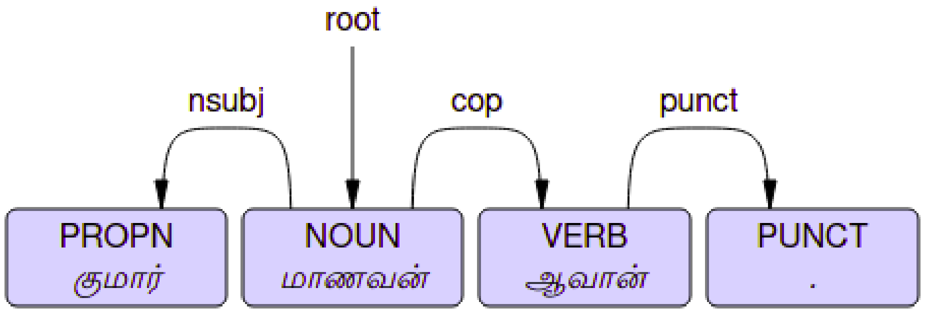
(37) குமார் வர மாட்டான் .



15. cop: copula

A cop (copula) is the relation of a function word used to link a subject to a nonverbal predicate

(38) குமார் மாணவன் ஆவான்.



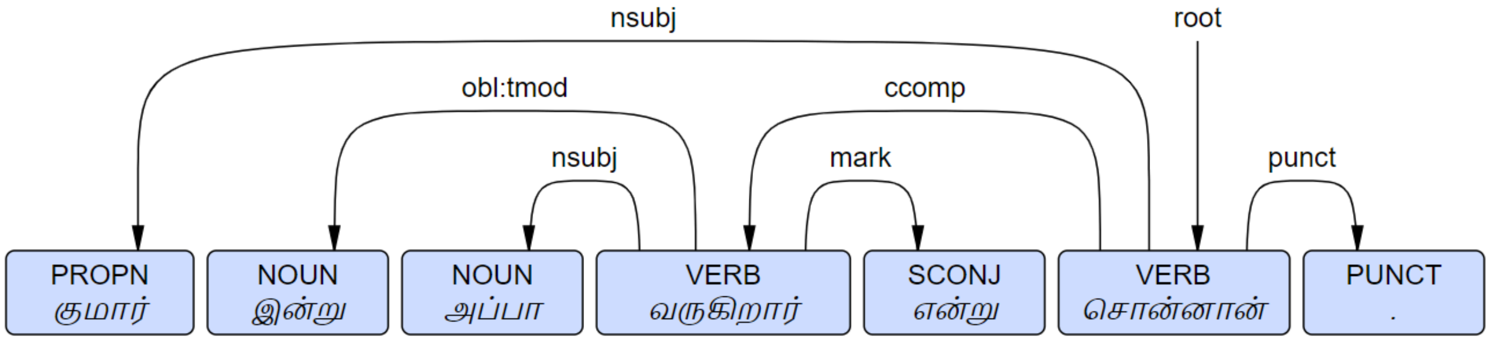
16. mark: marker

A marker is the word marking a clause as subordinate to another clause.

- For a complement clause, this is words like [en] *that* or *whether.*

- For an adverbial clause, the marker is typically a subordinating conjunction like [en] *while* or *although*.
- The marker is a dependent of the subordinate clause head.
- In a relative clause, it is a normally uninflected word, which simply introduces a relative clause, such as **என்று** (eṉṟu).

(39) குமார் இன்று அப்பா வருகிறார் என்று சொன்னான்.

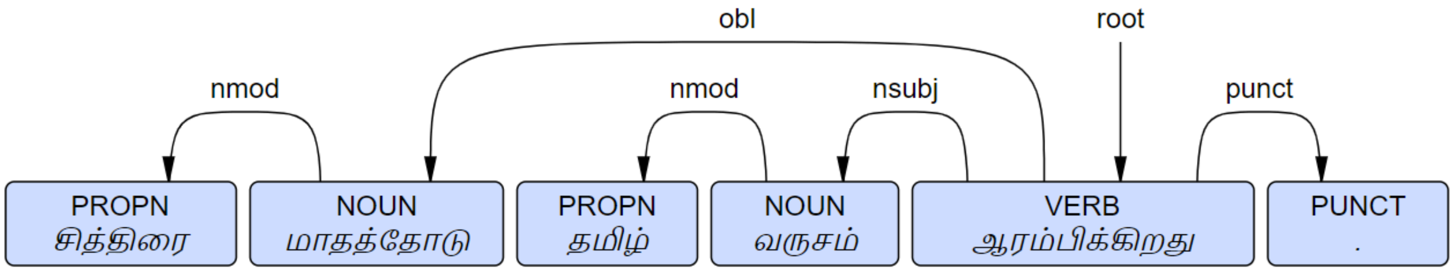


Nominal dependents

17. nmod: nominal modifier

The nmod relation is used for nominal dependents of another noun or noun phrase and functionally corresponds to an attribute, or genitive complement.

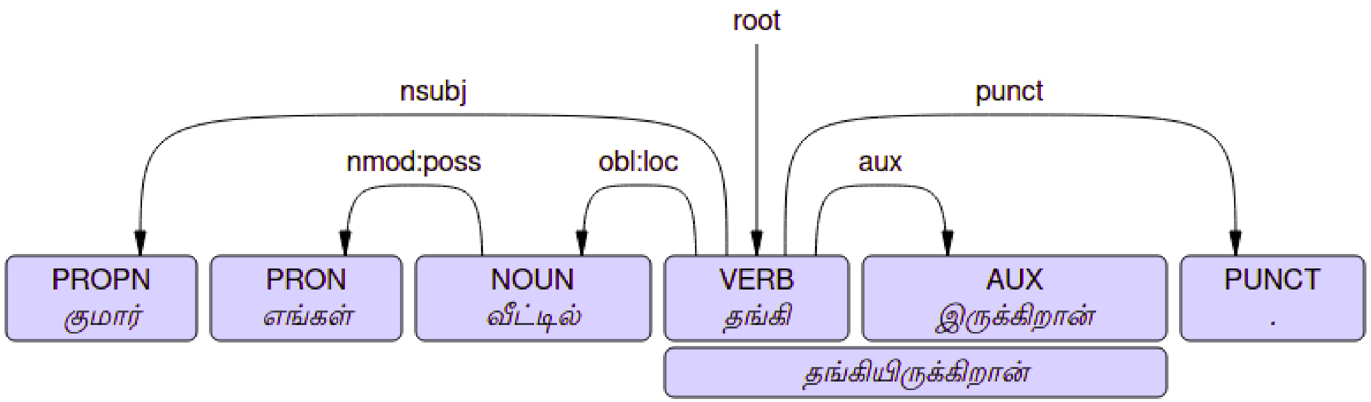
(40) சித்திரை மாதத்தோடு தமிழ் வருசம் ஆரம்பிக்கிறது .



17.1. nmod:poss

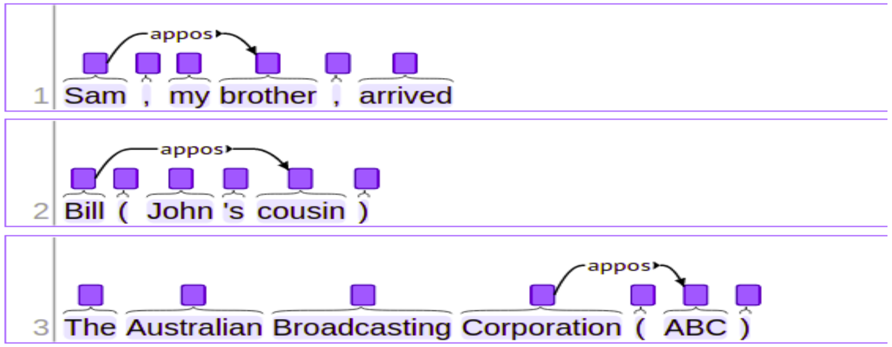
nmod:poss is used for a nominal modifier that occurs before its head in the specifier position in an oblique or possessive marker.

(41) குமார் எங்கள் வீட்டில் தங்கியிருக்கிறான் .



18. appos: appositional modifier

An appositional modifier of a noun is a nominal immediately following the first noun that serves to define, modify, name, or describe that noun. It includes parenthesized examples, as well as defining abbreviations in one of these structures.



1 Sam , my brother , arrived

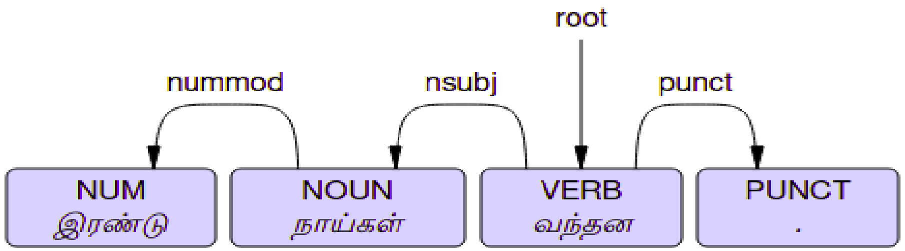
2 Bill (John 's cousin)

3 The Australian Broadcasting Corporation (ABC)

19. nummod: numeric modifier

A numeric modifier of a noun is any number phrase that serves to modify the meaning of the noun with a quantity.

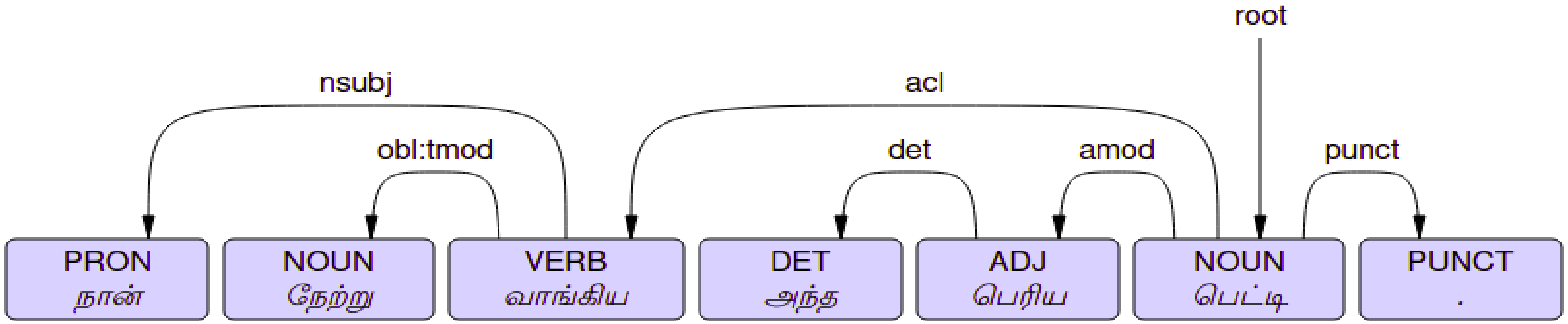
(42) இரண்டு நாய்கள் வந்தன .



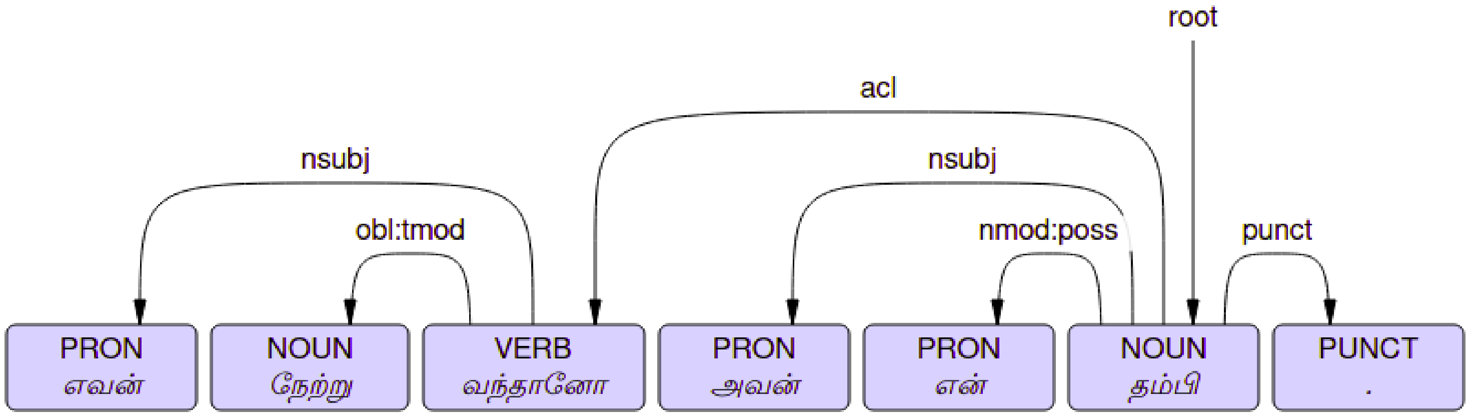
20. acl: clausal modifier of noun (adnominal clause)

acl stands for finite and non-finite clauses that modify a nominal. The acl relation contrasts with the advcl relation, which is used for adverbial clauses that modify a predicate. The head of the acl relation is the noun that is modified, and the dependent is the head of the clause that modifies the noun.

(43) நான் நேற்று வாங்கிய அந்த பெரிய பெட்டி .



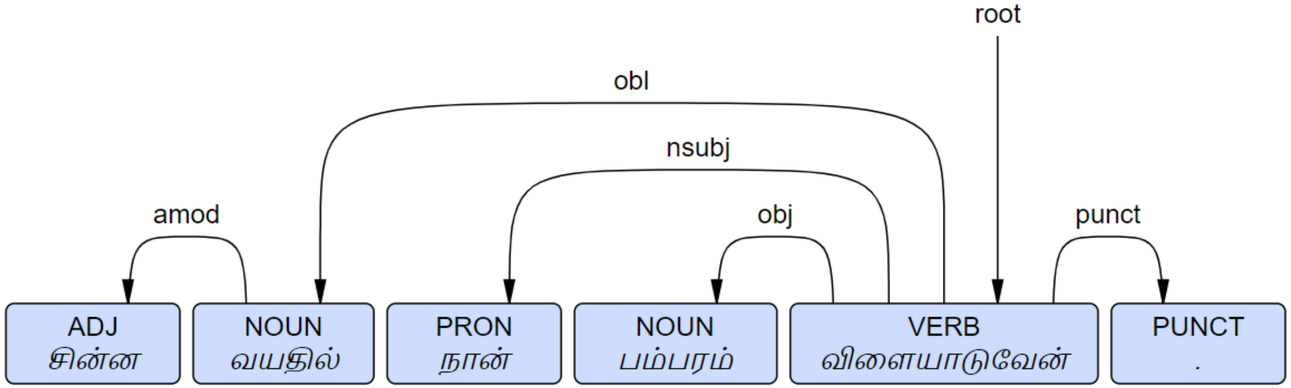
(44) எவன் நேற்று வந்தானோ அவன் என் தம்பி .



21. amod: adjectival modifier

An adjectival modifier of a noun (or pronoun) is any adjectival phrase that serves to modify the noun (or pronoun). The relation applies whether the meaning of the noun is modified in a compositional way (e.g., *large house*) or an idiomatic way (*hot dogs*).

(45) சின்ன வயதில் நான் பம்பரம் விளையாடுவேன் .

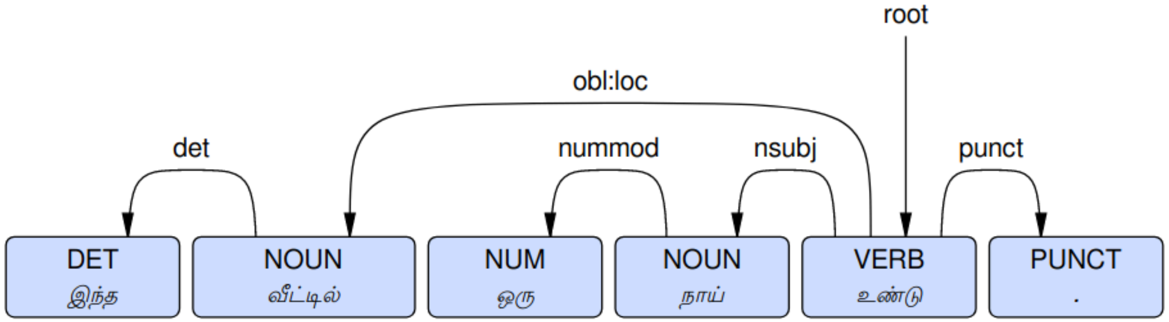


22. det: determiner

The relation determiner (det) holds between a nominal head and its determiner. Most commonly, a word of POS DET will have the relation det and vice versa.

Exceptions: In English, *my* is currently given the POS tag DET. But in Tamil, such possessive determiners are marked as nmod, so that it is parallel with other possessive constructions.

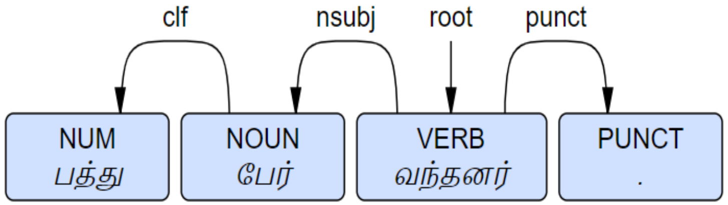
(46) இந்த வீட்டில் ஒரு நாய் உண்டு .



23. clf: classifier

A clf (classifier) is a word which accompanies a noun in certain grammatical contexts. The most canonical use is numeral classifiers, where the word is used with a number for counting objects.

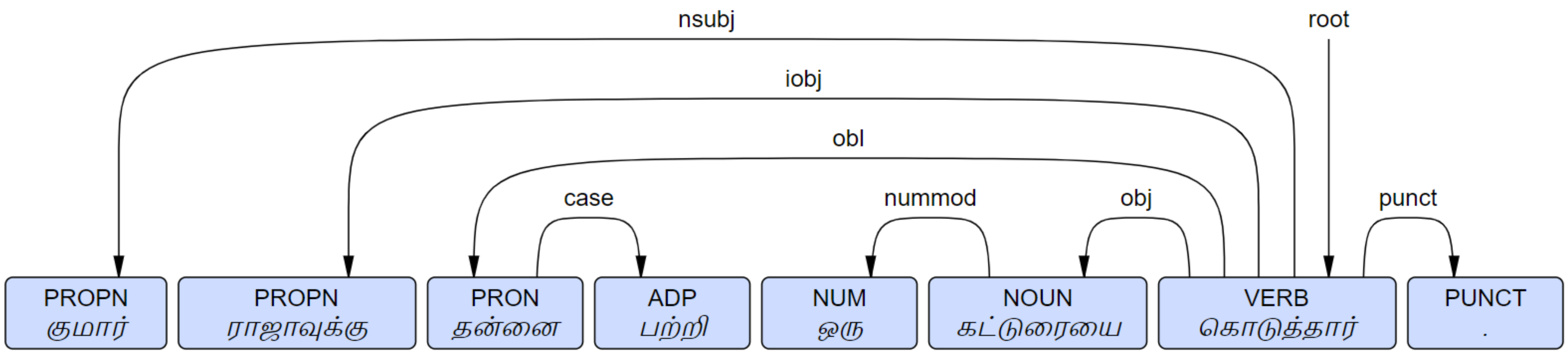
(46) பத்து பேர் வந்தனர்



24. case: case marking

The case relation is used for any case-marking element which is treated as a separate syntactic word (including prepositions, postpositions, and clitic case markers).

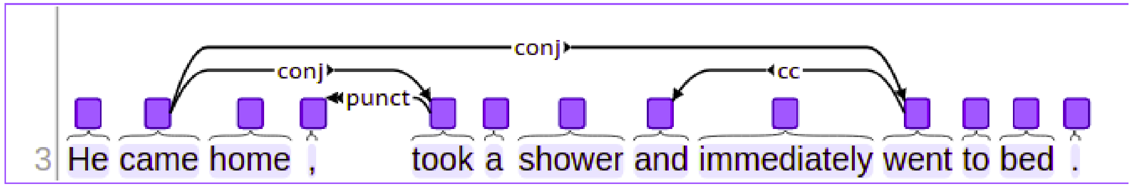
(47) குமார் ராஜாவுக்கு தன்னை பற்றி ஒரு கட்டுரையை கொடுத்தார் .

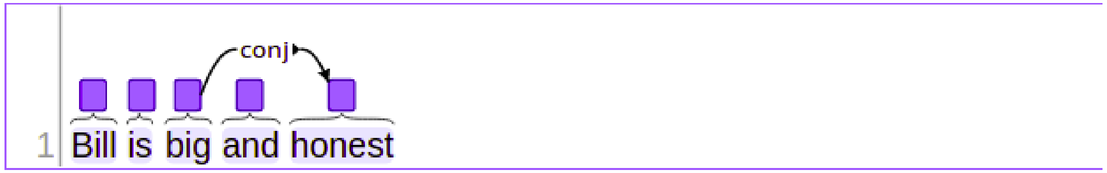


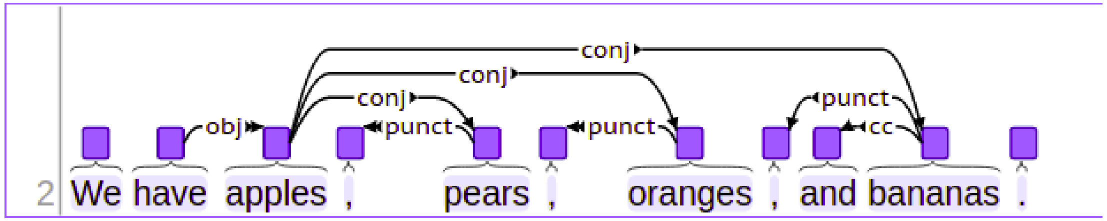
25. conj: conjunct

A conjunct is the relation between two elements connected by a coordinating conjunction, such as *and, or,* etc. Conjunctions are treated asymmetrically: The head of the relation is the first conjunct and all the other conjuncts depend on it via the conj relation.

Example:

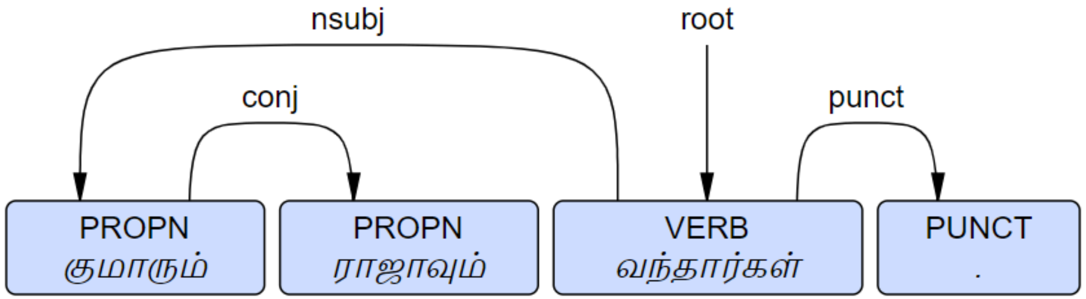

3 He came home , took a shower and immediately went to bed .


1 Bill is big and honest


2 We have apples , pears , oranges , and bananas .

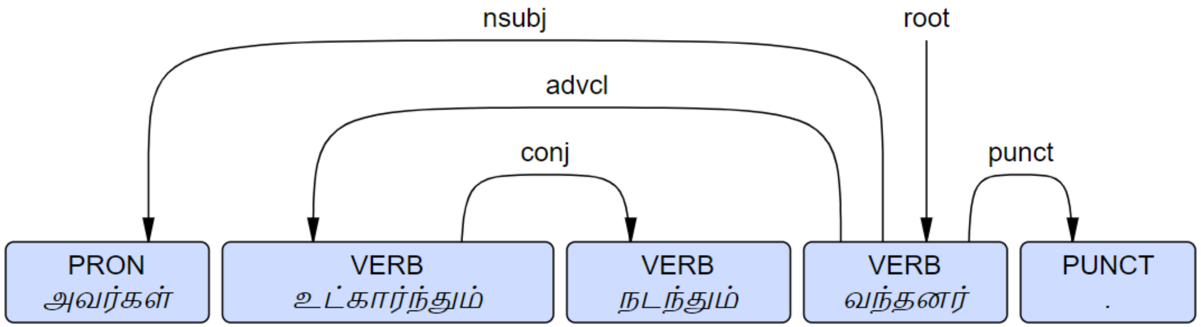
25.1. Noun-Noun conj

(48) குமாரும் ராஜாவும் வந்தார்கள் .



25.2. verb-verb conjunction

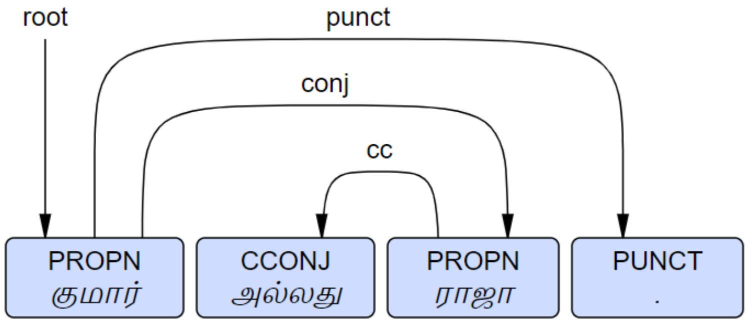
(49) அவர்கள் உட்கார்ந்தும் நடந்தும் வந்தனர்.



26. cc: coordinating conjunction

A cc is the relation between a conjunct and a preceding coordinating conjunction. In Tamil, ‘-um’ need not be parted.

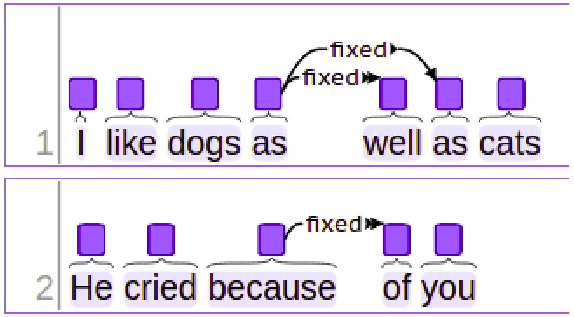
(50) குமார் அல்லது ராஜா .



27. fixed: fixed multiword expression

The fixed relation is one of the three relations for multiword expressions (MWEs) (the other two being flat and compound). It is used for certain fixed grammaticized expressions that behave like function words or short adverbials.

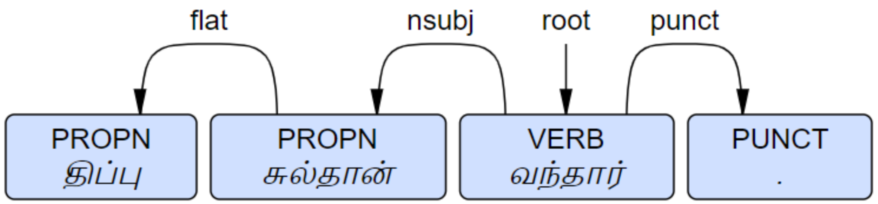
Example:



28. flat: flat multiword expression

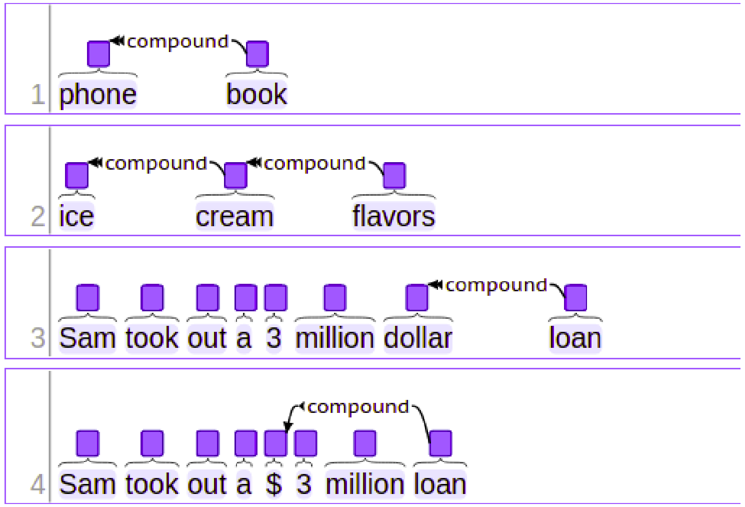
The flat relation is one of three relations for multiword expressions (MWEs) in UD (the other two being fixed and compound). It is used for exocentric (headless) semi-fixed MWEs like names (*Hillary Rodham Clinton*) and dates (*24 December*). It contrasts with fixed, which applies to completely fixed grammaticized (function word-like) MWEs (like *in spite of*), and with compound, which applies to endocentric (headed) MWEs (like *apple pie*).

(51) திப்பு சுல்தான் வந்தார் .



29. compound: compound

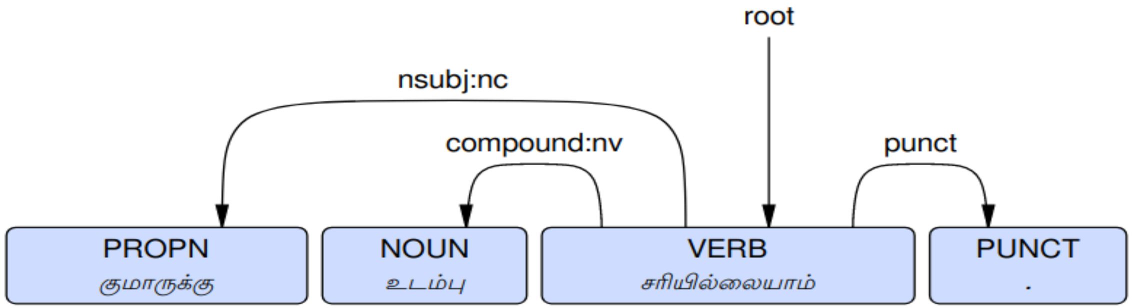
The compound relation is used for any kind of compounding: noun compounds (e.g., *phone book*), but also verb and adjective compounds that are more common in other languages (such as Persian or Japanese light verb constructions).



In Tamil, NV compounds are commonly seen.

29.1. compound:nv

(52) குமாருக்கு உடம்பு சரியில்லையாம் .

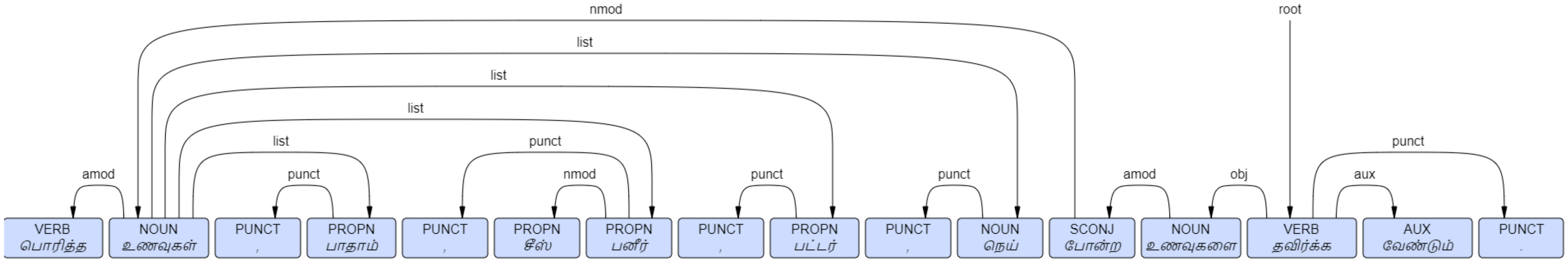


30. list: list

The list relation is used for chains of comparable items. In lists with more than two items, all items of the list should modify the first one. However, ‘list’ should not be overused. If a construction can easily be analysed using the grammatical relations of standard sentences, typically as a coordinated structure, then it should be analysed with these more standard relations, even if it is laid out as a list typographically. In particular, when the list is written as a single sentence, with commas and overt coordination, then it should be analysed as a coordinated structure.

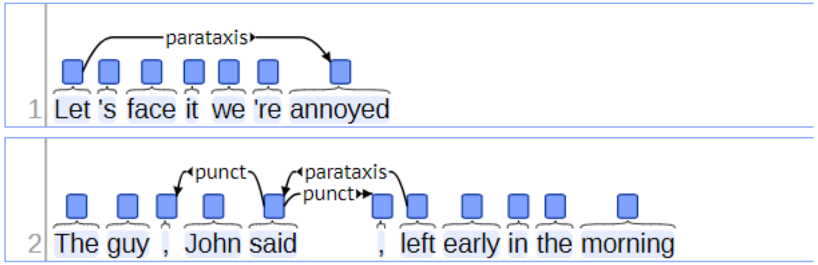
(53) பின்பு பொரித்த உணவுகள், பாதாம், சீஸ் பனீர், பட்டர், நெய் போன்ற உணவுகளை தவிர்க்க வேண்டும்.

Here, பாதாம், பனீர், பட்டர், நெய் are list to உணவுகள்.



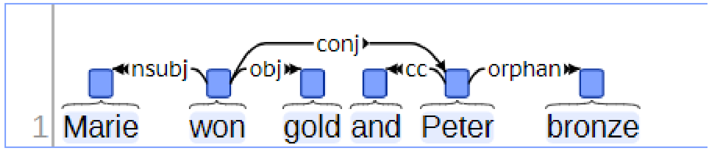
31. parataxis: parataxis

The parataxis relation (from Greek for “place side by side”) is a relation between a word (often the main predicate of a sentence) and other elements, such as a sentential parenthetical or a clause after a “:” or a “;”, placed side by side without any explicit coordination, subordination, or argument relation with the head word.


parataxis
1 Let 's face it we 're annoyed
punct
parataxis
punct
2 The guy , John said , left early in the morning

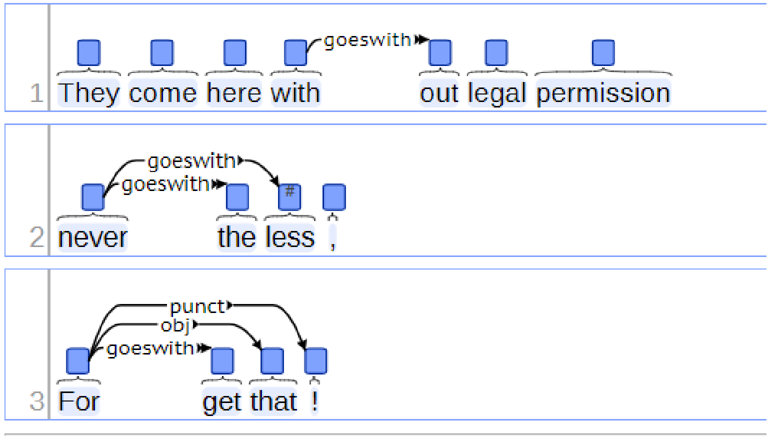
32. orphan: orphan

The ‘orphan’ relation is used in cases of head ellipsis where simple promotion would result in an unnatural and misleading dependency relation. The typical case is predicate ellipsis where one of the core arguments has to be promoted to clausal head.



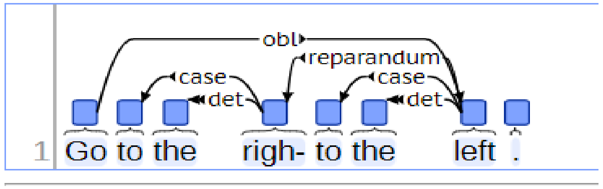
33. goeswith: goes with

This relation links two or more parts of a word that are separated in text that is not well edited. These parts should be written together as one word according to the orthographic rules of a given language. The head is always the first part, the other parts are attached to it with the ‘goeswith’ relation (for consistency, similarly as in flat, fixed and conj).



34. reparandum: overridden disfluency

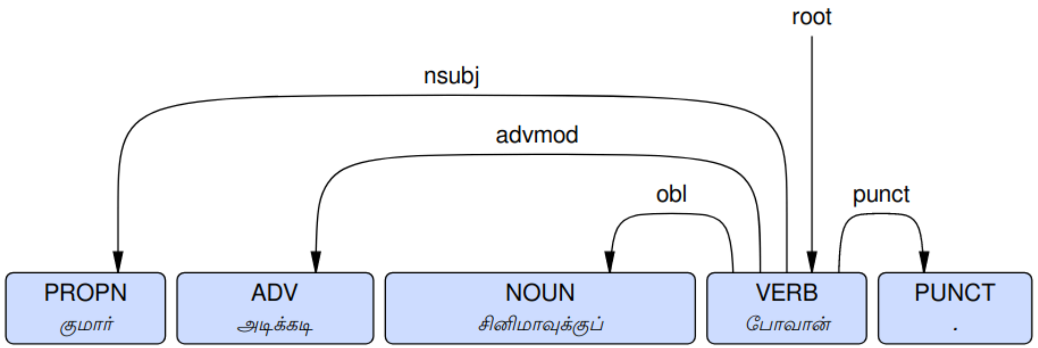
'reparandum' is used to indicate disfluencies overridden in a speech repair. The disfluency is the dependent of the repair. It is not found in the data as scientific and short story corpus is used for annotation.



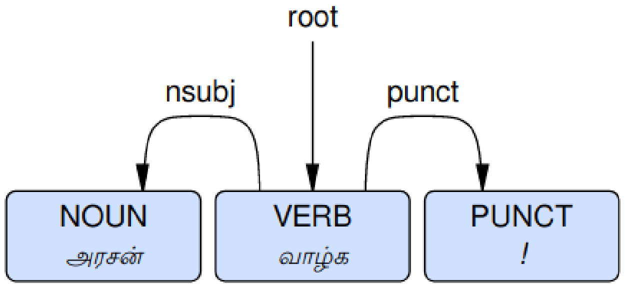
35. punct: punctuation

This is used for any piece of punctuation in a clause, if punctuation is being retained in the typed dependencies. Note that symbols tagged SYM are not punctuation and cannot be attached via the punct relation.

(54) குமார் அடிக்கடி சினிமாவுக்குப் போவான் .



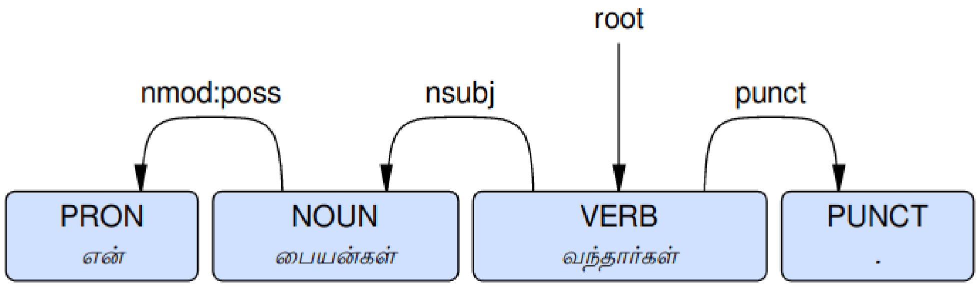
(55) அரசன் வாழ்க !



36. root: root

The root grammatical relation points to the root of the sentence. A fake node ROOT is used as the governor. In Tamil, verb or noun occur as 'root' in the majority of the sentences.

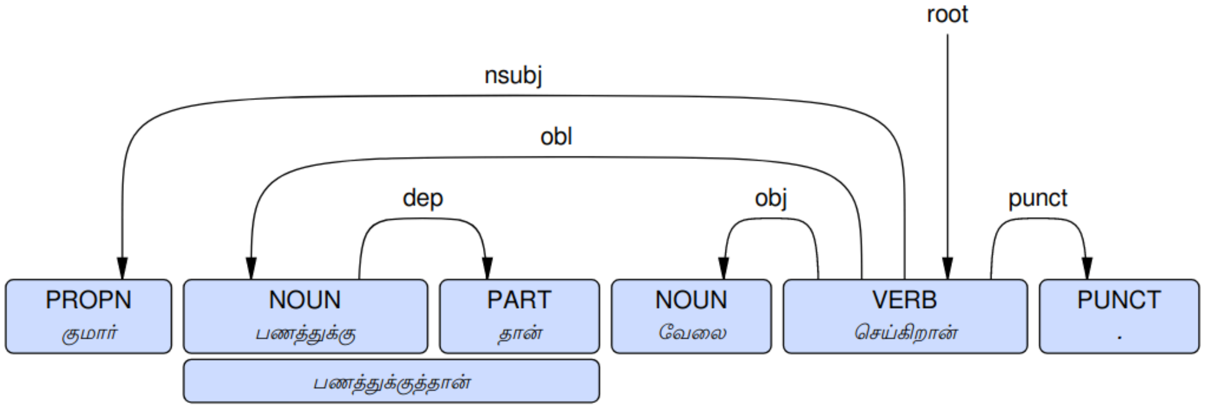
(56) என் பையன்கள் வந்தார்கள் .



37. dep: unspecified dependency

A dependency can be labeled as dep when it is impossible to determine a more precise relation. This may be because of a weird grammatical construction, or a limitation in conversion or parsing software. The use of dep should be avoided as much as possible.

(57) குமார் பணத்துக்குத்தான் வேலை செய்கிறான் .


root
nsubj
obl
dep
obj
punct
PROPN
குமார்
NOUN
பணத்துக்கு
PART
தான்
NOUN
வேலை
VERB
செய்கிறான்
PUNCT
.
பணத்துக்குத்தான்

One-to-one mapping:
ADJ-amod n+1
DET-det n+1
ADP-case n-1
AUX-aux n-1
ADV-advmod n+1; n-1

[இந்த நீர்நிலையில் ஹேங்கவுட் செய்வதற்கு] [ஒரு நிதானமான இடமாக இருப்பதை விட] [அதிகமான விஷயங்கள்_nsubj உள்ளன.]

NOUN/PRON/PROPN - subj (0 marker, Agreement with Verb-Gender, number and Person)

Clause- Verb (complete sentence/incomplete sentence)
Matrix clause-complete sentence
Subordinate clause- incomplete sentence

Phrase- NP, VP, ..

[நீங்கள் எங்கு சென்றாலும்,] [ஒரு கதை உங்களுக்குச் சொல்லப்படுகிறது.]

Nmod:clt,Nmod:clt:taan,nmod:clt:oo,nmod:clt:ee,Nmod:clt:aa,nmod:clt:um

----

Test for nsubj: \

(i) NOUN/ PRON/PROPN are subjects

(ii) NOUN , 0-marking

(iii) NOUN GNP (agreement) = VERB GNP

Test for nsubj:pass

(i) NOUN/ PRON are subjects

(ii) NOUN , 0-marking

(iii) NOUN GNP (agreement) = VERB GNP

(iv) Verb-படு

Active Voice Construction

(1) ரவி-0 கமலாவைப் பார்த்தான் (Ravi saw Kamala)

(a) Subject - 0 marking (Tag: nsubj)

(b) Object - ai marking (Tag: obj)

(c) Verb has no auxiliary, but GNP agreement with Subject

Passive Voice Construction

(2) ரவியால் கமலா-0 பார்க்கப்பட்டாள் (kamala was seen by ravi)

(a) Subject -0 marking (actually object) (Tag: nsubj:pass)

(b) ரவியால் **is Tag: (obl:agent)**

—------------

Test Obj

—-------------

- **NOUN/PROPN/PRON are objects**

- VERB should be **transitive**

1) Intransitive (I walk; I run; I sleep) , OBJECT IS NOT REQUIRED; Test: Passive cannot be done
2) Transitive (I saw him, I beat him…) , OBJECT IS REQUIRED; Test: Passive can be done
3) Ditransitive (I gave a book to him), obj, INDIRECT OBJECT (iobj); Test: iobj required ob

- **Object with -ai marking (Noun with (+animate) feature) e.g. person name, animate being; dog , horse etc..)**
  - ரவி கமலாவைப் பார்த்தான்
- Object with -0 marking ((**Noun with (-animate) feature) e.g. chair, table, cinema …)**

- ரவி படம் பார்த்தான்

—------------

Test iobj

—-------------

-NOUN/PROPN/PRON are objects

- VERB should be di**transitive**

-Object with -kku marking

ரவி புத்தகத்தை கமலாவுக்குக் கொடுத்தான்

1	ரவி	PROPN	4	nsubj
2	புத்தகத்தை	NOUN	4	obj
3	கமலாவுக்குக்	PROPN	4	iobj
4	கொடுத்தான்	VERB	0	root

நான் வந்தேன் . *நாம் வந்தேன்   (*-Ungrammatical)

வந்தேன் - Gend=any, Num=sg, Person=1

Output tree:

http://lindat.mff.cuni.cz/services/udpipe/

Tree Generation:

https://urd2.let.rug.nl/~kleiweg/conllu/