Guidlines for Chunk Types in Malayalam

A typical chunk consists of a single content word surrounded by a constillation of function words [S. Abney , 1991].

Malayalam being an agglutinative language, have a complex morphological and Syntactic structure.

Chunking is the task of identifying and segmenting the text into syntactically correlated word groups.

Chunking is the next step after POS Tagging, which divides sentences into non-recursive inseparable phrases. It is the task of identifying and labelling the simple phrases in a sentence from the tagged out put.

Chunk Types in Malayalam

A chunk would contain a head and its modifiers.

The following chunk types are seen in Malayalam.

1 NP - Noun Chunk

Noun chunk will be given the tag NP and include non-recursive noun phrases and post positional phrases. The head of the noun chunk would be a noun. Specifiers will come on the left side of the noun chunk and the vibhaktis will be part of the noun chunk. Particles, if any, will come on the right side of the noun chunk.

See these examples:

((വീടുകൾ))NP
houses

((വലിയ വീടുകൾ))NP
big houses

((വളരെ വലിയ വീടുകൾ))NP
very big houses

((സാരി)) NP
saree

((ഈ പുതിയ സാരി)) NP
this new saree

((ഈ പുതിയ സാരിയെ പറ്റി))NP

this new saree-Acc. about

((ഈ പുതിയ സാരിയെ പറ്റി പോലും))NP

this new saree- Acc. About even

These are all different types of chunks in Malayalam.

1.1 Chunk with a Genitive case

See the example belo

രാമന്റെ മകൻ
Rama-of son

'Rama's Son'

The above sentence will be chunked into 2 NPS ((രാമന്റെ)) NP and ((മകൾ)) NP , because it includes two nouns. But if the first one is a pronoun and the second one a noun, then these will come in one chunk

((രാമന്റെ)) NP

((മകൻ)) NP

((അവളുടെ മകൾ)) NP
she – of daughter

So, when Genitive suffix comes between 2 Hours, we must chunk it into two and when Genitive comes between a pronouns and a Noun, we must chunk together.

2. Verb Chunk

A verb group consists of a verb root, tense, mood, aspect and Negative, Question, Emphatic suffix etc:

Verbs can be active or passive, intransitive, transitive or causative.

Types of Verb Chunk

2.1 VGF - Finite Verb Chunk

In finite verb chunk, the verb group will be finite. The finiteness of the group is known by its Auxiliaries.

അമ്മ	ആഹാരം	കഴിച്ചു
mother	food	ate

' Mother ate the food'

In the above sentence, the action of te verb is complete and so it is a finite verb. The verb group which is finite is marked as VGF.

രാമൻ	പാടിക്കൊണ്ടിരിക്കുന്നു
Rama	sing-progr - prog- LM – pres.

' Rama is singing'.

The above verb group consists of the verb root പാട് - and the Auxiliaries-

കൊണ്ട് - , - ഇരി-, -ക്ക്-, and -ഉന്നു.

The following sentences are with finite verb.

സീത	പാട്ട്	പാടുന്നു
Sita	song	sing-pres.

'Sita sings the song'

അച്ഛൻ	നാളെ	വരും
Father	tomorrow	come-will

'Father will come tomorrow'.

2.2 VGNF - Non- finite Verb Chunk

The Adjectival participles, the Relative Participle form of verbs, the Adverbial Participle forms, the conditional forms, the Adverbials denoting time, the Nominalized Gerunds, the Verbal Participles etc: are included in VGNF chunk.

2.2.1 Adjectival participle form

ഡൽഹിയിൽ താമസിക്കുന്ന എന്റെ സഹോദരൻ

'Delhi-in' 'staying' ' my' 'brother'
നാളെ വരും.
'tomorrow' ' come-will'

' My brother who is staying in Delhi will come tomorrow.'

2.2.2 Relative Participle form

ആ പാടുന്ന പെൺകുട്ടി രാജന്റെ മകൾ ആണ്.
That sing-RP girl Rajan-of daughter is

'That girl who is singing is Rajan's daughter'.

Here, പാടുന്ന 'who sings' is the Relative participle form and so it is chunked as VGNF.

2.2.3 Adverbial Participle form

രാമൻ ഡ്രൈവറായി ജോലിയിൽ പ്രവേശിച്ചു
Rama Driver-as Job-in entered

'Rama entered job as a Driver'

In this sentence, ഡ്രൈവറായി 'as Driver' is marked as VGNF.

2.2.4 Conditional forms

നീ വന്നാൽ നമുക്ക് സിനിമയ്ക്ക് പോകാം.
you came-if we-to film-to go-shall

' If you come, we shall go for film'

Here, വന്നാൽ is conditional form and it is marked as VGNF.

2.2.5 Adverbials denoting time

ഞാൻ വന്നപ്പോൾ അദ്ദേഹം ഉറങ്ങുകയായിരുന്നു
I come- when he (Hon.) sleep-was

'when I came, he was sleeping'

Here, വന്നപ്പോൾ is an Adverbial form denoting time and so it will be marked as VGNF.

2.2.6 Adverbial Participle form - അവെ 'while'

ഞാൻ അവിടെ നില്ക്കവെ രാമനെ കണ്ടു
I there stand-while Ram-Acc. Saw

'While I was standing there, I saw Ram'

Here, നില്ക്കവെ 'while standing ' is marked as VGNF.

All these forms (1-6) are marked as VGNF because they are derived from a verb and can have their arguments.This information is useful for processing at the syntactic level.

2.2.7 Nominalized verb forms

Nominalized verb forms with gender suffixes like വന്നവൻ 'he who came', അഭിനയിച്ചവൾ she who acted, പറഞ്ഞത് 'that which is said' will be chunked as VGNF because they are capable of taking their Arguments.

ഇന്നലെ വന്നവൻ രാമൻ ആണ്
Yesterday came-NOL Ram is
'The man who came yesterday is Ram'

In the above sentence, വന്നവൻ 'he who came' is marked as VGNF.

100 പടത്തിൽ അഭിനയിച്ചവൾ ആണ് സീത
100 film-in acted-NOML is Sita
' Sita is the one who acted in 100 films'

Here, അഭിനയിച്ചവൾ' She who acted' is marked as VGNF.

അത് പറഞ്ഞത് രാമൻ ആണ്

that said-NOML Ram is

'It is Ram who said that'

In the sentence, പറഞ്ഞത് is the Nominalized form and അത് പറഞ്ഞത് is a clause and the whole clause is Karma Karaka (K2). It is marked as VGNF because it can take arguments.

Eventhough these Nominalized forms with gender suffixes-അൻ,-അൾ and -ത് are labelled as VGNF, they have the Nominal qualities of taking case suffixes and postpositions. See below:

വന്നവനെ 'he who came-Acc.,' വന്നവന് 'he who came-to', വന്നവനെ കൊണ്ട് 'he who came (Acc.)-with',അഭിനയിച്ചവളോട് 'she who acted-with',അഭിനയിച്ചവളുടെ കൂടെ 'along with she who acted', പറഞ്ഞതിനെ 'that which is said-Acc.', പറഞ്ഞതിനെ പറ്റി 'that which is said-Acc.-about' etc.

The above said fact is important and notable one. The above said fact is a good evidence to include all these Nominalized forms among Nouns.

Ignoring all these facts, if we split these Nominalized forms into RP form and Pronoun, then the sentence will not be acceptable to the native speaker of Malayalam.

For example, take the sentence below:

അത് പറഞ്ഞത് രാമൻ ആണ്.

That said-NOML Rama is

'It is Rama who said that'

If we split it, it will be as seen below:

അത് പറഞ്ഞ അത് രാമൻ ആണ്

that which is said that Rama is

We can see that the resulting sentence is unacceptable and ungramatical.

From the above said facts,it is clear that we cannot ignore the truth of the nominal quality of the above said peculiar forms. It is a peculiarity of Malayalam. In day today life, we Malayalees use many sentences with Nominalized verbs.

2.2.8 Verbal Participle

രാമൻ മരത്തിൽ ചാടി കയറി
Rama tree- in having jumped climbed

'Having jumbed, Rama climbed the tree'

Here, ചാടി 'having jumped' is Verbal participle and it is marked as VGNF.

ഞാൻ അവിടെ പോയിട്ട് അത് സംസാരിച്ചു.
I there having gone that discussed

'Having gone there, I discussed that'
Here, പോയിട്ട് ' having gone' is VGNF.

2.3 VGINF - Infinitive Verb Chunk

This tag is to mark the infinitival verb forms. In Malayalam, the Purposive Infinitive is an Infinitive verb chunk.

See below:
അമ്മ കുളിക്കാൻ പോയി
mother take bath-to went
'Mother went to take bath'

In the above sentence, കുളിക്കാൻ 'to take bath' is a Purposive Infinitive form and it is labelled as VGINF.

രാവിലെ കുളിക്കുക പ്രയാസം ആണ്
morning to take bath difficult is

'To take bath in the morning is difficult'

In the above sentence, കുളിക്കുക 'to take bath' is infinitive and annotated as VGINF.

2.4 VGNN - Verbal Nouns

Nouns derived from verbs are Verbal Nouns. These verbal Nouns will be annotated as VGNN.

See the example, below:

പാടത്ത് ഇന്ന് കൊയ്ത്ത് ഉണ്ട്.
farm-in today harvest is

'There is harvest in the farm today'

In the above sentence, കൊയ്ത്ത് 'harvest' is VGNN. It is derived from the verb കൊയ്യുക 'to harvest'.

See another sentence,

ഞാൻ അവളുടെ ഓട്ടം കണ്ടു
I She-of running saw
'I saw her running'

Here, ഓട്ടം 'running' is VGNN, and it is derived from the verb ഓടുക 'to run'.

3. JJP - Adjectival Chunk

An adjectival chunk will be tagged as JJP. This chunk includes all adjectives,including predicative adjectives.

See this sentence:

ആ പെൺകുട്ടി((വളരെ സുന്ദരി_JJ))JJP ആണ്
that girl very beautiful is

Adjectives appearing before a noun will be grouped together with the noun chunk.

((നല്ല_JJ ചുവന്ന_JJ പുതിയ_JJ സാരി_NN))NP
good red new saree

Here, the three adjectives come before the noun and so these will be grouped together with the noun chunk.

4. RBP - Adverb chunk

An Adverb occuring separately will be marked as RBP.

കള്ളൻ പതുക്കെ പതുക്കെ ജനലിന്റെ അരികിലേയ്ക്ക് വന്നു
thief slowly slowly window-of near-towards came

'The thief slowly slowly came near the window'

Now examine the following sentence:

അവർ രസിച്ചുല്ലസിച്ച് നടക്കുകയായിരുന്നു
they having delighted and thrilled walking- were

'Having delighted and thrilled, they were walking.'

അച്ഛൻ ആഹാരം കഴിച്ചിട്ട് ഉറങ്ങി.
father food after eating slept

' Father slept after eating his meal.'

In the above examples, ആർത്തുല്ലസിച്ച് and കഴിച്ചിട്ട് are nonfinite forms of Verbs which are used as Adverbs. Similar to Adjectival Participles, these will also be chunked as VGNF and not as RBP. The reason is that we need to preserve the information that these are underlying verbs. This will be a crucial information at the Dependency level where the Arguments of these verbs will also be marked.

5. NEGP - Negative Particle

There is no separate Negative Particle chunk in Malayalam. Negatives are suffixes in Malayalam.

6. CCP - Conjuncts

Conjuncts are functional units about which is required to build the larger structures.

6.1 Co-ordinating Conjuncts

Co-ordinating Conjuncts are not free forms in Malayalam. See below:

രാമൻ പഠിക്കുകയും സീത ഉറങ്ങുകയും ചെയ്യുന്നു
Rama study-and Sita sleep-and doing

'Rama is studying and Sita is sleeping.'

Here, -ഉം is co-ordinating suffix and it is not a separate chunk in Malayalam.

6.2 Subordinating conjuncts - CCP

രാമൻ നാളെ വരും എന്ന് സീത പറഞ്ഞു
Rama tomorrow come-will that Sita said

'Sita said that Ram will come tomorrow'

In the above sentence, എന്ന 'that' is CCP in Malayalam.

ജോൺ മരിച്ചു എന്ന വാർത്ത ശരി ആണ്
John died that news true is

'The news that John died is true.'

Here, എന്ന 'that' is CCP.

The 2 sentences above will be chunked as follows:

((രാമൻ))_NP ((നാളെ വരും))_VGF ((എന്ന്))_CCP ((സീത))_NP ((പറഞ്ഞു))_VGF

((ജോൺ))_NP ((മരിച്ചു))_VGF ((എന്ന))_CCP
((വാർത്ത))_NP ((ശരി))_NP ((ആണ്))_VGF

7. FRAGP - Chunk Fragment

There are no chunk Fragments in Malayalam.

8. BLK - Miscellaneous entities

Entities such as interjections and discourse markers, that cannot fall into any of the above mentioned chunks will be kept with in a separate chunk (BLK).

((ഓ_INJ))BLK

((ഹാവൂ_INJ))BLK

9. Some special cases

9.1 Conjunct Verbs

The Noun/Adjective and verb (internal component of a conjunct verb) will be chunked separately in a conjunct verb.

പോലീസ് അയാളെ ചോദ്യം ചെയ്തു
police he-Acc. question done

' Police questioned him'

In the above example, ചോദ്യം ചെയ്തു 'questioned' is a conjunct verb. It will be chunked as shown below:

((ചോദ്യം))_NP ((ചെയ്തു))_VGF

9.2 Particles

Particles will be chunked with the same chunk as the anchor word they occur with. See below:

((രാമനും കൂടി))_NP 'Rama also'

((രാമൻ പോലും))_NP 'even Rama'

9.3 Quantifiers

Numbers can occur either as noun modifiers before a noun or can occur without a noun,with inflections.

eg: ആയിരം കുട്ടികൾ '1000 students'

ആയിരങ്ങളുടെ 'of thousands'

These will be chunked as follwos:

((ആയിരം_QC കുട്ടികൾ_NN))_NP
'thousands' 'students'

((ആയിരങ്ങളുടെ))_NP
'of thousands'

A QC or QO occuring with a noun will be part of the noun chunk. All categories occuring without a noun, with nominal inflections, will be tagged as noun.

9.4 Punctuations

All punctuations, with an exception of sentence boundary markers and clausal conjuncts, will be included in the preceding chunk. See the example below:

അവൻ പറഞ്ഞു,"ഇത് ശരി അല്ല"

((അവൻ_PRP))_NP ((പറഞ്ഞു_VF_SYM))_VGF

(("_SYM ഇത്_PRP))_NP ((ശരി_JJ))_JJP

((അല്ല_VM"_SYM))_VGF

"He said, " it is not true"

Punctuations such as hyphens and quote marks will be taken care of by the to kanizer.