Chunking

Chunking

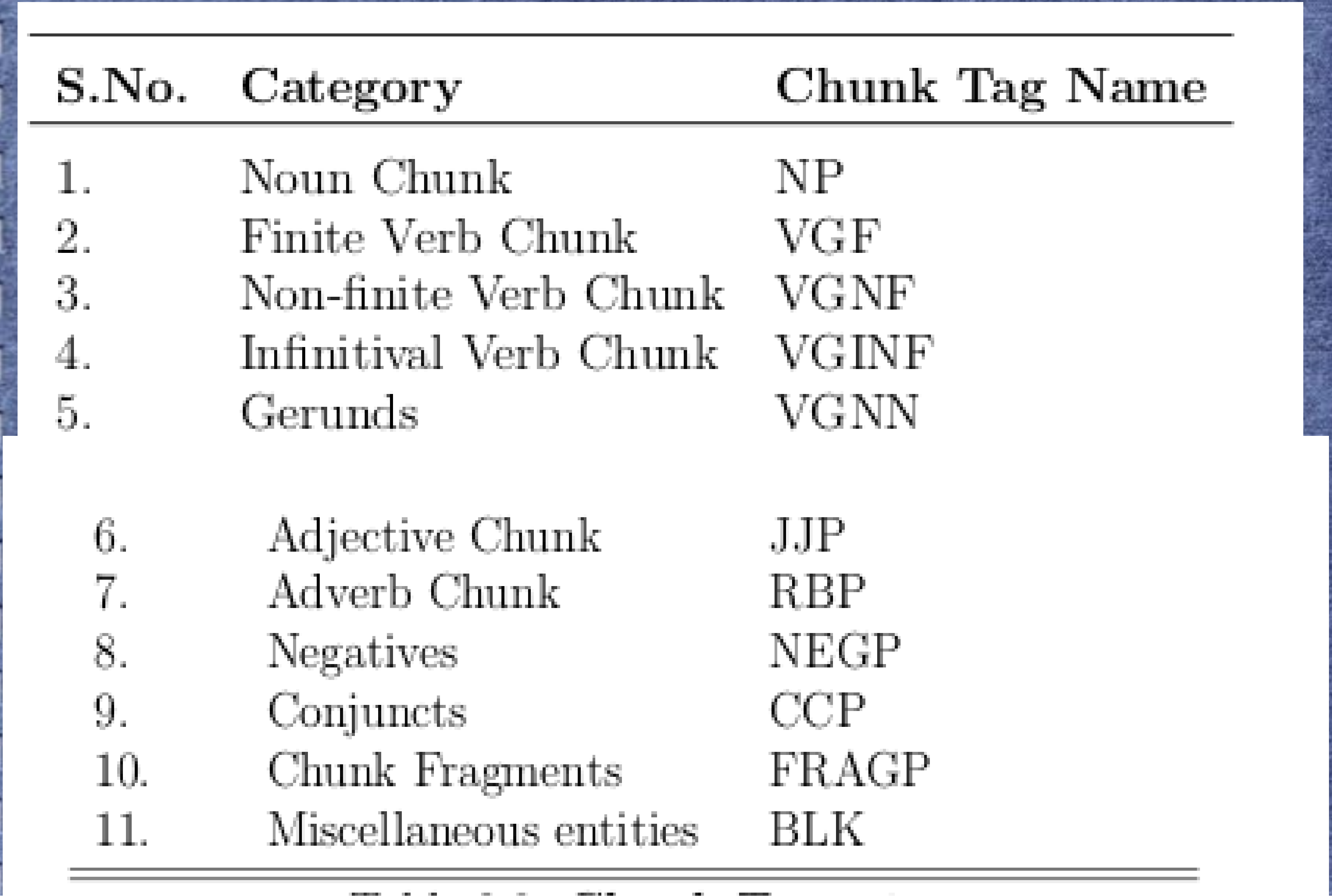
Chunking involves in identification and tagging of the non-recursive combinations of word groups.

It groups **syntactically related immediate sequences of words**.

A typical chunk consists of a single content word surrounded by a constellation of function words (Abney,1991).

Chunks are normally taken to be a '**correlated group of words**'.

Chunk Types



S.No.	Category	Chunk Tag Name
1.	Noun Chunk	NP
2.	Finite Verb Chunk	VGF
3.	Non-finite Verb Chunk	VGNF
4.	Infinitival Verb Chunk	VGINF
5.	Gerunds	VGNN
6.	Adjective Chunk	JJP
7.	Adverb Chunk	RBP
8.	Negatives	NEGP
9.	Conjuncts	CCP
10.	Chunk Fragments	FRAGP
11.	Miscellaneous entities	BLK

1. Noun Chunk - NP

1. (*bacce*_NN)) _NP

'children'

2. ((*kucha*_QF *bacce*_NN))_NP

'some' 'children'

3. ((*kucha*_QF *acche*_JJ *bacce*_NN))_NP

'some' 'good' 'children'

4. ((*Dibbe*_NN *meM*_PSP))_NP,

'box' 'in'

5. ((eka_QC *kAlA*__JJ *ghoDZA*_NN))_NP ,

'one' 'black' 'horse'

6. ((*yaha*_DEM *nayI*_JJ *kitAba*_NN))_NP,

'this' 'new' 'book'

7. ((**isa**_DEM *nayI*_JJ *kitAba*_NN
*meM*_PREP))_NP,

'this' 'new' 'book' 'in'

8. ((*isa*_DEM *nayI*_JJ *kitAba*_NN *meM*_PSP
*bhI*_RP))_NP

'this' 'new' 'book' 'in'
'also'

The issue of genitive marker and its grouping with the nouns.

'*rAma kA beTA*'

((*rAma kA*))_NP, ((*beTA*))_NP

Therefore, it was decided that the genetive markers will be chunked along with the preceding noun. Thus, the noun group 'rAma kA beTA' would be chunked into two chunks.

((*rAma kA*))NP ((*beTA*))NP acchA hE "Ram's son is good"

((*kitAba*))NP **((*rAma kI*))NP** ((hE)) VGF "The book belongs to Ram"

For the noun groups such as "*usakA beTA*" it was decided that they should be chunked together.

2. Verb Chunks

The verb chunks would be of several types. A verb group will include the main verb and its auxiliaries, if any. Following are some examples of verb chunks from Hindi,

((*khAyA*)), ((*khA rahA hE*)), ((*khA sakawe hEM*))

'ate' 'eat' 'PROG' 'is' 'eat' 'can' 'PRES'

VGF - Finite Verb Chunk

mEMne ghara para khAnA **((*khAyA*_VM))_VGF**

'I erg' 'home' 'at' 'meal' 'ate'

vaha cAvala

((*khA*_VM *rahA*_VAUX *hE*_VAUX))_VGF

'he' 'rice' 'eat' 'PROG' 'is'

VGNF Non-finite Verb Chunk

A non-finite verb chunk will be tagged as VGNF. For example,

> *seba* ((***khAtA***_**VM** ***huA***_**VAUX))_VGNF**

laDZakA jA rahA thA

'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

> *mEMne ghAsa* ((***khAte***_**VM** ***hue***_**VAUX))_VGNF** *ghoDe ko dekhA*

'I erg' 'grass' 'eating' 'PROG' 'horse' acc 'saw'

VGINF Infinitival Verb Chunk

This tag is to mark the infinitival verb form.

In Hindi, both, gerunds and infinitive forms of the verb end with a -*nA* suffix.

Since both behave functionally in a similar manner, the distinction is not very clear.

However, languages such as Bangla etc have two different forms for the two types. Examples from Bangla are given below.

Borabela ***((snAna karA))*****_VGNN** *SorIrera pokze BAlo*

'Morning' 'bath' 'do-verbal noun' 'health-gen' 'for' 'good'

‘Taking bath in the early morning is good for health”

bindu Borabela *((****snAna karawe))*****_VGINF** *BAlobAse*

'Bindu' 'morning' 'bath' 'take-inf' 'love-3pr'

“Bindu likes to take bath in the early morning”

VGNN Gerunds

A verb chunk having a gerund will be annotated as VGNN. For example,

> *sharAba* ((**pInA**_VM))_VGNN *sehata ke liye hAnikAraka hE.*

liquor' 'drinking' 'heath' 'for' 'harmful' 'is'

“Drinking (liquor) is bad for health”

> *mujhe rAta meM* ((**khAnA**_VM))_VGNN *acchA lagatA hai* 'to me' 'night' 'in' 'eating' 'good' 'appeals'

“I like eating at night”

h20a. ((**sunane_VM** **meM_PSP))_VGNN** *saba kuccha acchA ((lagatA_VM hE_VAUX))_VGF*

'listening' 'in' 'all'
'things' 'good' 'appeal' 'is'

JJP Adjectival Chunk

An adjectival chunk will be tagged as JJP. This chunk will consist of all adjectival chunks including the predicative adjectives. However, adjectives appearing before a noun will be grouped together with the noun chunk. A JJP will consist of phrases like

vaha laDaZkI hE((*suMdara*_**JJ** *sI*_RP))_**JJP**

'she' 'girl' 'is' 'beautiful' 'kind of'

hAthI *AyA* ((*moTA*_*C *tagadA*_**JJ**))_**JJP**

'elephant' 'came' 'fat' 'powerful'

vaha laDakI ((*bahuta*_INTF *sundara*_**JJ**))_**JJP** *hE*

'she' 'girl' 'very' 'beautiful' 'is'

Cases such as (h61) below will not have a separate JJP chunk. In such cases, the adjectives will be grouped together with the noun they modify. Thus forming a NP chunk.

h61. ((*kAle*_**JJ** *ghane*_**JJ** *laMbe*_**JJ** *bAla*_**NN**))_**NP**

'black' 'thick' 'long' 'hair'

Following examples from Hindi present a

h62. *xillI meM* ***rahanevAlA*** *merA BAI kala A rahA hE .*

'Delhi' 'in' 'staying' 'my' 'brother' 'tomorrow' 'come' 'PROG' 'is'

“My brother who stays in Delhi is coming tomorrow”.

h63. *usane Tebala para* ***rakhA huA*** *seba khAyA.*

'(s)he erg' 'table' 'on' 'kept' 'apple' 'ate'

“He ate the apple kept on the table”.

In (h62) above '*rahanevAlA*' is an adjectival participle. But we do NOT mark it as JJP. Instead, it will be marked as a **VGNF**. The decision to tag it as a VGNF is based on the fact that such adjectival participles are derived from a verb can have their arguments. This information is useful for processing at the syntactic level. Thus, '*rahanevAlA*' in (h62) will be annotated as follows:

h62a. *xillI*   *meM* ((***rahanevAlA*_VM)_VGNF** *merA BAI kala A rahA   hE* .

Similarly, in (h63) above, the chunk '*rakhA huA*' is an adjective but will also be marked as a VGNF since this also derived from a verb and chunks like '*Tebala pra*' etc are its arguments. So the chunk name will be **VGNF** and the POS tag will be **VM** which might be followed by an auxiliary verb tagged as **VAUX**. (h63a) shows how '*rakhA huA*' will be annotated :

h63a. *usane Tebala para ((* **rakhA_VM huA_VAUX))_VGNF***seba khAyA.*

RBP Adverb Chunk

This chunk name is again in accordance with the tags used for POS tagging. This chunk will include all pure adverbial phrases.

h64. *vaha* ((*dhIre_RB dhIre_***RB**))_**RBP** *cala rahA thA.*

'he' 'slwoly' 'walk' 'PROG' 'was'

“He was walking slowly”

*vaha ((dhIre_RB))_***RBP** *cala rahA thA.*

Now consider the following examples:

h65. *vaha* ***dagamagAte hue*** *cala rahA thA .*

'he'   '   'walk' 'PROG' 'was'

“He was walking

h66. *vaha khAnA* ***khAkara*** *ghara gayA .*

'he'   'meal'   'after eating' 'home' 'went'

“He went home after eating his meal”

In the above examples, '*dagamagAte hue*' and 'k*hAkara*' are non finite forms of verbs used as adverbs. Similar to adjectival participles these will also be chunked as **VGNF** and not as **RBP**. The reason for this is that we need to preserve the information that these are underlying verbs.

NEGP Negatives

In case a negative particle occurs around a verb, it is to be grouped within verb group. For example,

h67. *mEM kala* *dillI* *((***nahIM*****_NEG** ***jA*****_VM** ***rahI*****_VAUX*****))*****_VGF**
"I" "tomorrow" "Delhi" "not" "go" "Cont"

h68. ((***binA*****_NEG** ***bole*****_VM))_VGNF** *kAma* *((***nahIM*****_NEG** ***calatA*****_VM*****))*****_VGF**
"without" "saying" "work" "not" "happen"

However ,

h69. **binA** kucha **bole** kAma **nahIM calatA**

"without" "something" "saying" "work" "not" "happen"

In the above sentence, the noun "*kucha*" is coming between the negative "*binA*' and verb "*bole*".

Here, it is not possible to group the negative and the verb as one chunk. At the same time, "*binA*" cannot be grouped within an NP chunk, as functionally, it is negating the verb and not the noun.

To handle such cases an additional **NEGP** chunk is introduced.

If a negative occurs away from the verb chunk, the negative will be chunked by itself and chunk will be tagged as NEGP. Thus,

h69a. **((*binA*))_NEGP** **((**_kucha_))_NP **((*bole*))_VG**
((_kAma_))_NP ((_nahIM calatA_**))_VG**

CCP Conjuncts

Conjuncts are functional units information about which is required to build the larger structures. Take the following examples of cunjunct usages :

h70. (*rAma kitAba paDha rahA thA)* **Ora** *(mohana Tennisa khela rahA thA)*.

“Ram was reading a book **and** Mohan was playing tennis”

h71. (*rAma ne batAyA)* ***ki*** *(usakI kitAba acchI hE)*.

“Ram said **that** his book is good”

h72. (*rAma)* **Ora** *(mohana) Tennisa khela rahe the*.

“Ram **and** Mohan were playing tennis”.

h70a. ((*rAma*))_NP ((*kitAba*))_NP ((*paDha rahA thA*))_VG **((*Ora*))CCP** ((*mohana*))_NP ((*Tennisa*))_NP ((*khela rahA thA*))_VGF.

h71a. ((*rAma ne*))_NP ((*batAyA*))_NP **((*ki*))_CCP** ((*usakI*))_NP ((*kitAba*))_NP ((*acchI*))_JJP ((*hE*))_VGF.

FRAGP Chunk Fragments

Some times certain fragments of chunks are separated from the chunks to which they belong. For example :

h76. ***rAma*** (*jo* *merA baDZA bhAI* *hE*) ***ne*** *kahA* ...

'Ram' 'who' 'my' 'elder' 'brother' 'is' 'erg' 'said'

In the above example, vibhakti '*ne*', which is a case marker of the noun '*rAma*', is separated from it by an intervening clause. Syntactically, '*ne*' is a part of the noun chunk '*rAma ne*'. However, at times it can be written separately. The following was decided for such fragments :

There will be a separate chunk for the vibhakti in constructions where it gets separated from the noun it would normally be grouped with. This chunk can have more than one entity within it.

h77. ((***rAma***))**_NP,** *mere dillI vAle bhAI,* **((*ne*))_FRAGP** *kahA*

'Ram' 'my' 'Delhi' 'from' 'brother' 'erg' 'said'

(ii) If the entities embedded between the noun and it's vibhakti are a series of nouns the entire group will be chunked as a single noun chunk.

h78. **((***isa* ***'upanyAsa samrATa'*** *Sabda kA***))_NP**

'this' 'Novel' 'King' 'word' 'of'

BLK Miscellaneous entities

Entities such as interjections and discourse markers that cannot fall into any of the above mentioned chunks will be kept within a separate chunk.

eg. ((*oh*_INJ))_**BLK,** ((*arre*_INJ))_**BLK**