POS Tagset for Malayalam

Part-of-speech (POS) tagging, also known as grammatical tagging, is the process of marking the words in a text as corresponding to a particular part of speech, based on both its definition and context. It is the commonest form of the corpus annotation. Falls under Syntactic processing stage in NLP : One of the preliminary tasks in language processing.

Let's understand with a simple example:

കേരളം വിട്ട് ലോകത്തിന്റെ ഏത് കോണിൽ പോയാലും ബന്ധുക്കളോടും മറ്റും സംസാരിക്കുവാനും വിവരങ്ങൾ അന്വേഷിക്കുവാനും വേണ്ടി മാതൃഭാഷ അറിഞ്ഞിരിക്കേണ്ടത് അത്യാവശ്യം ആണ്.

After tagging, we get

കേരളം	വിട്ട്	ലോകത്തിന്റെ	ഏത്	കോണിൽ	പോയാലും	ബന്ധുക്കളോടും	മറ്റും	സംസാരിക്കുവാനും	വിവരങ്ങൾ
N_NNP	**V_VM_VNF**	**N_NN**	**PR_PRP**	**N_NN**	**V_VM_VNF**	**N_NN**	**RP_RPD**	**V_VM_VNF**	**N_NN**

അന്വേഷിക്കുവാനും	വേണ്ടി	മാതൃഭാഷ	അറിഞ്ഞിരിക്കേണ്ടത്	അത്യാവശ്യം	ആണ്	.
V_VM_VNF	**PSP**	**N_NN**	**V_VM_VNF**	**N_NN**	**V_VM_VF**	**RD_PUNC**

The tagset for Malayalam is as follows

Sl.No	Category			Label	Annotation Convention	Examples
	Top Level	Sub Type Level (1)	Sub Type Level (2)			
1	Noun			N	N	മോഹൻ, വീട്, 1989-ൽ
1.1		Common		NN	N_NN	വീട്, വെള്ളം, പട്ടം
1.2		Proper		NNP	N_NNP	മോഹൻ, രവി, കൊല്ലം, 1989-ൽ, ചിങ്ങം, നയാഗ്ര, താജ്മഹൽ
1.3		Noun Locative		NST	N_NST	മുകളിൽ, താഴെ, മുന്നിൽ, പിന്നിൽ, ഇടയിൽ
2	Pronoun			PR	PR	അവിടെ, ഇപ്പോൾ, അന്ന്
2.1		Personal		PRP	PR_PRP	ഞാൻ, നീ, അവ, നമ്മുടെ, അവൻ, അവൾ, അത്, ഇത്, ഞങ്ങൾ, നിങ്ങൾ
2.2		Reflexive		PRF	PR_PRF	താൻ, തന്നെത്താൻ, സ്വയം
2.3		Relative		PRL	PR_PRL	ആരോ, എന്തും

2.4		Reciprocal		PRC	PR_PRC	തമ്മിൽ, തമ്മിൽതമ്മിൽ, പരസ്പരം
2.5		Wh-word		PRQ	PR_PRQ	ആര്, എന്ത്
3	Demonstrative			DM	DM	ആ, ഈ
		Deictic		DMD	DM_DMD	അത്, ഇത്
		Relative		DMR	DM_DMR	ഏത്
		Wh-word		DMQ	DM_DMQ	എങ്ങനെ, ഏത്
4	Verb			V	V	പോയി, നടക്കുന്നു, ആണ്, ഉണ്ട്
4.1		Main		VM	V_VM	പോ-, കഴി-, ചിരി-, ആണ്
4.1.1			Finite	VM	V_VM_VF	പോയി, ചിരിക്കും, കഴിക്കുന്നു, ആണ്
4.1.2			Non-finite	VNF	V_VM_VNF	പോയ, ചിരിച്ച, കഴിച്ച, പറഞ്ഞത്, വന്നത്, ഉള്ള
4.1.3			Infinitive	VINF	V_VM_VINF	വരാൻ (വരുവാൻ), പറയുക, കഴിക്ക്, ചിരിച്ചാൽ

4.2		Verbal Noun		VN	V_VN	പഠിത്തം, നടത്തം, നെയ്ത്ത്, ഓട്ടം, വിൽക്കൽ
5	Adjective			JJ	JJ	വലിയ, ചെറിയ, പല, ചില, നീല, അഴകുള്ള, സുന്ദരമായ
6	Adverb			RB	RB	വേഗം, പതുക്കെ, പെട്ടെന്ന്, പതിവായി, ഒട്ടാകെ, മുമ്പോട്ട്, എതിരെ
7	Post position			PSP		പറ്റി, കൂടെ, കുറിച്ച്, ഒപ്പം
8	Conjunctions			CC	CC	പക്ഷേ, എന്നിട്ടും, എന്നാൽ, എന്നാലും, എങ്കിലും
8.1		Co-ordinator		CCD	CC_CCD	പക്ഷേ, അഥവാ, അല്ലെങ്കിൽ
8.2		Subordinator		CCS	CC_CCS	എന്നാൽ, എന്നിട്ടും, അതുപോലെതന്നെ, എങ്കിലും
8.2.1			Quotative	UT	CC_CCS_UT	എന്ന്, എന്ന
9	Particles			RP	RP	മാത്രം, കൂടി, തന്നെ
9.1		Default		RPD	RP_RPD	മാത്രം, കൂടി, പോലും, വെറും

9.2		Classifier		Ch	RP_CL	തുടങ്ങിയ, മുതലായ, എന്നീ, പേർ
9.3		Interjection		INJ	RP_INJ	അയ്യോ, ഹാവൂ
9.4		Intensifier		INTF	RP_INTF	വളരെ, ഏറെ, ഏറ്റവും, അത്യന്തം
10	Quantifier			QT	QT	കുറച്ച്, ധാരാളം, രണ്ട്
10.1		General		QTF	QT_QTF	കുറച്ച്, ധാരാളം, അധികം
10.2		Cardinals		QTC	QT_QTC	ഒന്ന്, രണ്ട്, രണ്ടിരട്ടി, അര, **10**
10.3		Ordinals		QTO	QT_QTO	ഒന്നാം, രണ്ടാം, ഒന്നാമത്തെ, **10%** രണ്ടാമത്തെ, **1-**ാം, **3-**ാമത്തെ
11	Residuals			RD	RD	
11.1		Foreign word		RDF	RD-RDF	Any word written in a script other than the script of the original text. In this case, all words written in other languages like English, Hindi, Sanskrit, Tamil, etc
11.2		Symbol		SYM	RD-SYM	$, &, *,+,@,^, ...

11.3		Punctuation		PUNC	RD_PUNC	. , , , “ , " ?
11.4		Unknown		UNK	RD_UNK	Words not identified by the tagger
11.5		Echo-words		ECH	RD_ECH	ചടപടാ, ചറപറ