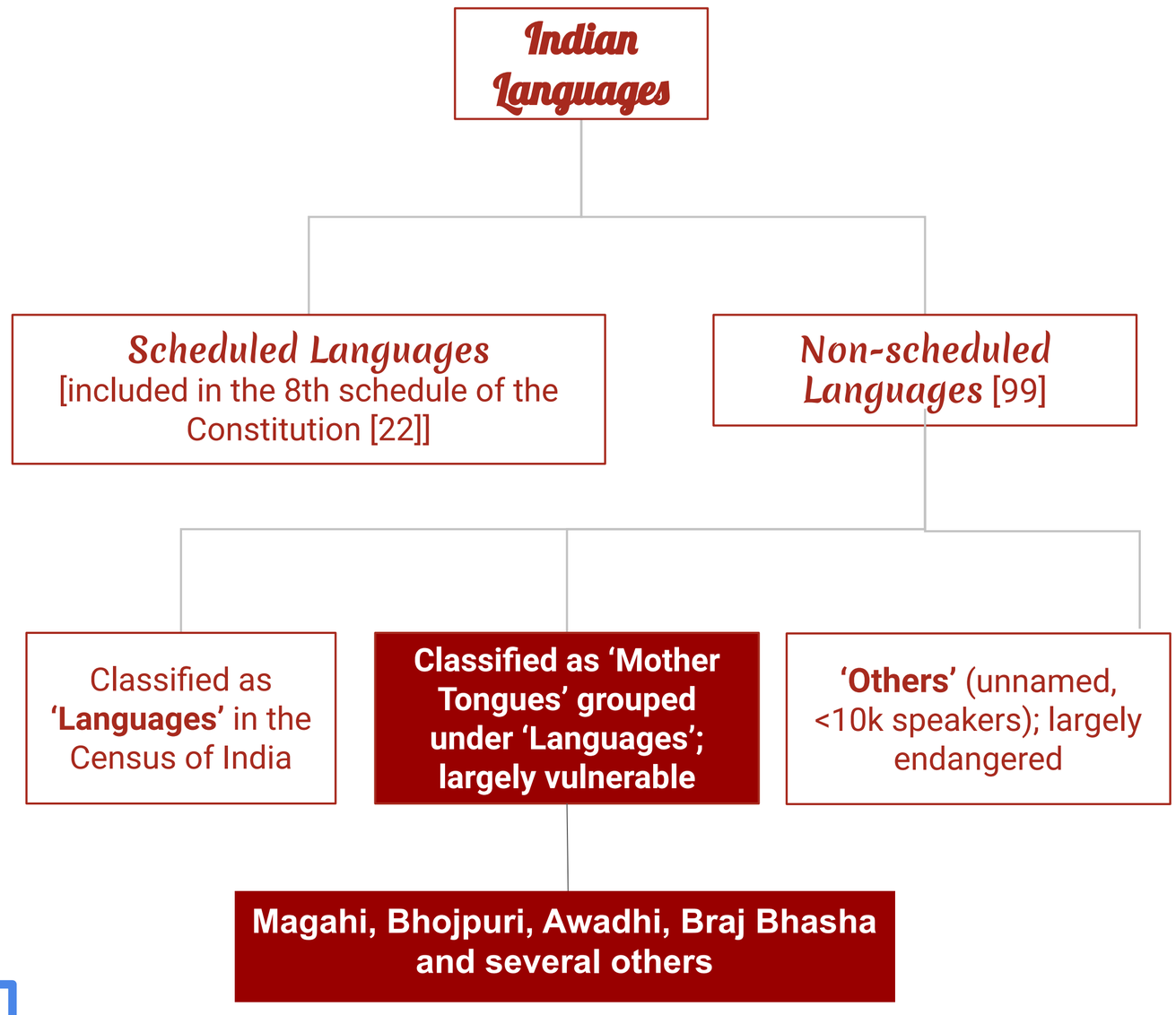
Language Resources and Technology Development Efforts for some Lesser-known Indian Languages

Ritesh Kumar, *KMI, Dr. Bhimrao Ambedkar University, India*; **Atul Kr. Ojha**, *Panlingua Language Processing LLP, India*
Bornini Lahiri, *IIT-Kharagpur, India*; **Mayank**, *Kendriya Hindi Sansthan, India*; **Deepak Alok**, *Rutgers University, USA*
All present and past **students of KMI, Agra** and **JNU, New Delhi** involved in these projects

Abstract

For the last few years, we have been involved in the development of language technologies and resources for some of the lesser-known Indian languages viz. Magahi, Bhojpuri, Awadhi and Braj Bhasha. These languages (among others) have been largely marginalised and ignored and have low prestige and negative attitude towards them because of these being considered 'illiterate' and 'rural' varieties of Hindi. Our poster will showcase different kinds of corpora as well as basic technologies like part-of-speech tagger, morphological analyser as well as some applications like machine translation systems that have been developed for these languages.

पिछला कै साल से, हमनी मगही, भोजपुरी, अवधी आउ ब्रज भासा जईसन भासा, जेकरा पर बहुते कम काम होल हई, ओकर प्रौद्योगिकि आउ संसाधन के विकास में लगल ही। इ सब भासा (औ अइसन बड़ीमनी) के नजर अंदाज और हासिया पर कर देल गेल हे आउ कम परतिस्ठा आउ नीचा दृस्टि से देखल जा हे काहे की इ सब हिंदी भासा के देहाती आउ असीछित बोली के प्रकार समझल जा हे। हमनी के पोस्टर बिभिन्न प्रकार के कॉर्पोरा के साथ-साथ बुनयादी प्रौद्योगिकि जैसे की सब्द भेद टैगर (पार्ट ऑफ़ स्पीच टैगर), रूप सम्बन्धी बिश्लेसक (मोरफोलॉजिकल अनलाइज़र) आउ कुछ अनुप्रयोग सॉफ्टवेयर जैसे की मसीनी अनुवाद पद्धति/सिस्टम जे इसब भासा ला विकिसित कइल गइल हे ओकरा परदरसीत करत।



Magahi Resources and Technologies

>200k sentences	Raw Corpus	• Collection of Folk Tales • Blogs • Modern Fiction & Non-fiction
55k sentences	POS-annotated Corpus	• Annotated with UD [manual and automatically]
10k sentences	Parallel Corpus	• English - Magahi • Russian - Magahi [non-aligned; auto created]
1.3k sentences	Morph-annotated Corpus	• Annotated with UD Feature Sets
1k sentences	UD Treebank	• Annotated with UD relations
0.99 / 0.94 F-score	POS Tagger	• SVM - based POS tagger
0.94 F-score	Morph Analyser	• Rule-based system • Partially implemented
	MAG - EN MT	• Phrase-based MT System

Bhojpuri Resources and Technologies

>100k sentences	Raw Corpus	• News and Blogs • Collection of Folk Tales • Modern Fiction and Non-fiction
65k sentences	POS-annotated Corpus	• Annotated with UD as well as BIS POS tags
65k sentences	Parallel Corpus	• English - Bhojpuri • Hindi-Bhojpuri (>65k)
4k tokens	UD Treebank	• Annotated with UD Tagset (including Morph, POS and relations)
86.78%	POS Tagger	• CRF - based POS tagger
	Eng↔Bho MT	• Phrase-based MT • Factor-based MT • Neural MT • Tree-based MT

Braj Bhasha Resources and Technologies

9k sentences	Raw Corpus	• Modern Fiction and Non-fiction
9k sentences	POS-annotated Corpus	• Annotated with BIS POS Tagset
1.6k sentences	Morph-annotated Corpus	• Annotated with UD Morph Features
88.27%	POS Tagger	• SVM-based POS Tagger

Awadhi Resources and Technologies

70k tokens	Raw Corpus	• Fiction and Stories
20k tokens	POS-annotated Corpus	• Annotated with UD POS Tagset
0.88 F-score	POS Tagger	• SVM based POS tagged

FOR COLLABORATION AND QUERIES:
panlingua@outlook.com