Issues and Challenges in Developing Statistical POS Taggers for Sambalpuri

Pitambar Behera[1(✉)], Atul Kr. Ojha[2], and Girish Nath Jha[1,2]

[1] Centre for Linguistics, Jawaharlal Nehru University, New Delhi, India
pitambarbehera2@gmail.com, girishjha@gmail.com
[2] Special Centre for Sanskrit Studies,
Jawaharlal Nehru University, New Delhi, India
shashwatup9k@gmail.com

Abstract. Low-density languages are also known as lesser-known, poorly-described, less-resourced, minority or less-computerized language because they have fewer resources available. Collection and annotation of a voluminous corpus for the purpose of NLP application for these languages prove to be quite challenging. For the development of any NLP application for a low-density language, one needs to have an annotated corpus and a standard scheme for annotation. Because of their non-standard usage in text and other linguistic nuances, they pose significant challenges that are of linguistic and technical in nature. The present paper highlights some of the underlying issues and challenges in developing statistical POS taggers applying SVM and CRF++ for Sambalpuri, a less-resourced Eastern Indo-Aryan language. A corpus of approximately 121 k is collected from the web and converted into Unicode encoding. The whole corpus is annotated under the BIS (Bureau of Indian Standards) annotation scheme devised for Odia under the ILCI (Indian Languages Corpora Initiative) Project. Both the taggers are trained and tested with approximately 80 k and 13 k respectively. The SVM tagger provides 83% accuracy while the CRF++ has 71.56% which is less in comparison to the former.

Keywords: Low-density language · Parts of speech tagger · SVM
CRF++ · Sambalpuri · Eastern IA Language

1 Introduction

Low-density languages have fewer resources in terms of the availability of voluminous corpus [1] for NLP applications. The unavailability of corpus for a low-density language proves to bear adverse impacts on its future NLP development. As rightly pointed out by Ostler [2], languages that lack active participation in the electronic media are doomed to be endangered in the forthcoming years. Most of these languages are either dialects or languages with no government recognition. As a result, the situations of these languages in South Asia in general and in Indic languages, in particular, are 'relatively bleak' [1]. Although India is a land of more than 6000 languages with five prominent diverse language families [3] only 22 are scheduled and the rest are fighting for their survival.

Z. Vetulani et al. (Eds.): LTC 2015, LNAI 10930, pp. 393–406, 2018.
https://doi.org/10.1007/978-3-319-93782-3_28

This paper is concerned with demonstrating the issues and challenges in developing statistical parts of speech taggers for a low-resource language nomenclatured as Sambalpuri or Kosli. The paper has broadly three objectives. Firstly, it highlights the issues in corpus collection with regard to non-uniform orthographic language standards and non-Unicode encodings of the written text. Secondly, it further attempts to bring out the issues in annotation having without any guideline. Finally, it demonstrates the challenges in developing statistical POS taggers for Sambalpuri owing to the typical, unobserved and language-specific linguistic nuances.

2 Sambalpuri: A Low-Density Eastern IA Language

Sambalpuri (ISO 639-3) is an Eastern Indo-Aryan (EIA) language and is also known as Dom, Kosli, Koshal, Koshali, Western Odia[1] etc. It is spoken in the ten districts of western and south-western Odisha which comprise Bargarh, Bolangir, Kalahandi, Sonepur, Sambalpur, Jharsuguda, Sundargarh, Deogarh, Boudh, Nuapada and Athmallik sub-division of Angul district. In comparison to its sister languages such as Maithili, Awadhi, Bhojpuri, Magahi, Angika, Bengali, Assamese, Odia and many others, Sambalpuri has not gained much attention; neither from linguists nor from the government. It is really quite obvious to affirm that it shares the genetic affinity with the Indo-Aryan language family by observing some of its linguistic features [4]. Although it has 70–76%[2] lexical similarity with Standard Oriya [5], it is syntactically a distinct language [6]. Recently, Behera and Dash [7] have reported a work-in-progress Sambalpuri dictionary which has the present size of around 500 lexicons pertaining to nouns in three domains such as flora and fauna, kinship and body parts.

3 Salient Linguistic Features

Less-described languages have some of the most interesting linguistic features that are typical and language-specific. Some of the features like agglutination, classifiers, serial verbs, multi-words, compounds etc. account for the less accuracy of any of the statistical NLP applications in general. Some of them are discussed vividly below in four sub-sections with regard to Sambalpuri: agglutination, classifiers, reduplication and compounds.

3.1 Agglutination

In an agglutinative language, words are made of a linear sequence of distinct morphemes each of which corresponds to a definite meaning[3]. In Odia, the categories such as "suffixes, postpositions, and case endings agglutinate with the verbs, nouns, adverbs or pronouns" [8–13]. Similar is the case with Sambalpuri language. Behera [14] has

[1] https://www.ethnologue.com/language/spv.

[2] However there is no satisfactory explanation about the methodology adopted and the number of lexical items analysed on the basis of which this conclusion has been arrived at.

[3] http://www.glossary.sil.org/term/agglutinative-language.

recently averred the fact that Sambalpuri has some agglutinative nominal morphology and assumes that it has some Dravidian features as Odia possesses.

For instance,
kʰɑɛbɑr-ke ‘to eat’
lʊk-ər ‘people’s’
bɑhɑr-ke ‘to outside’
mɒr-nʊ ‘from me’

In the examples instantiated above, all the case endings or markers /ke/, /ər/, /nʊ/ agglutinate with their head categories verb, noun and pronoun. /ke/ which is equivalent to the English infinitive and preposition ‘to’ is alternating here with both verb and spatial noun /kʰɑɛbɑr/ and /bɑhɑr/ respectively.

3.2 Classifiers

Classifier is one of the most prominent phenomena in Indian languages; especially in the Tibeto-Burman languages and Dravidian languages. In addition to some of the EIA languages like Bengali [15], Odia [16] and [10, 11], Bhojpuri [17], Marathi [18] and so on, it is a dominant linguistic feature in Sambalpuri as well. The classifiers mainly occur either as proper classifiers, attached to numerals or to the quantity word /keṇe/ ‘how many, some’, or as indefinite markers, in combination with the suffix” [16] as /ʈe/, /ʈɑ/, /kʰə̃ɖe/, /ɟʰəne/ etc. in Sambalpuri. One of the rarely observed phenomena of Indian languages found in Sambalpuri is that classifiers also occur with post positions. Recently, it has been asserted that classifiers alternation with other categories takes place largely in Sambalpuri [14].

For instance, /mɒr lekʰe-ʈɑ/ ‘like me’.

3.3 Reduplication

Reduplication is the repetition of a syllable, segment or some part or whole of a lexical or phrasal unit which leads to a semantic or grammatical modification of the component in question. There are two types of reduplication: partial and total. In total reduplication, the whole part of the base is reduplicated and in the partial reduplication, some part is reduplicated [19]. In the following instances, the first two are fully reduplicated while the rest of the following are partial. In the partial reduplication, the final syllables /-nɑ/ of both the words are reduplicated like in the first example whereas the final example contains the reduplications of the initial syllables /hʊ-/.

For instance,

çɪk çɪk ‘shining’
ɖʰɪre ɖʰɪre ‘slowly’
ɟənɑ sʊnɑ ‘known’
hʊʈɑ hʊʈɪ ‘abusing’

When a sequence of verbs occurs in a chronology, they are called serial verbs [20] and some of them are reduplicative in nature. In the below-instantiated example, it is quite confusing as to how to annotate the verbal occurrences. Because the initial verb

/d̪əʊɽ/ is a non-finite verb followed by a verbal reduplication which is functioning like a manner adverb modifying the finite following verb /pəleɪgəla/. The issue here is how to annotate the verbal reduplication. In these instances the morphological features of each lexical item have been taken into consideration for deciding the annotation of label.

For example,

se mɑrɪkərɪ d̪əʊɽ d̪əʊɽ pəleɪgəlɑ
he V-Nonfinite V-reduplication V-Finite
"He went away beating."

3.4 Compounds

Compound or Sandhi is one of the most productive linguistic phenomena which is quite typical in most of the worlds' languages in general and in Indian languages in particular. There are three basic types of compounds: vowel, consonant and visarga. In the following instance, the first word is an adjective and the second is a noun, but when get combined they comprise a nominal category. Since Sambalpuri is a head final language, the annotation label is decided on the basis of the category of the head. Here the head is a nominal element and hence the judgment goes in favor of the category of the label of the head word.

For example,

sɔd̪\JJ + pɔʈ̪^{h}ɔ\N_NN = sɔʈ̪pɔʈ̪^{h}ɔ\N_NN 'good path'

So, in the above example, the decision whether to annotate the word as an adjective or noun goes for the right-headedness feature of Sambalpuri. This feature is typical to most of the IA languages and the word /sɔʈ̪pɔʈ̪^{h}ɔ/ is labelled as a noun.

4 Methodology

This section deals with (a) the total corpus collected in four major domains, (b) the BIS annotation guideline adapted for Sambalpuri, (c) size of the corpus for training, testing and development stages and (d) features selection for SVM and CRF++ POS taggers.

4.1 Corpus Size

The tabulated data (see Table 1) demonstrates the total corpus size collected for developing the Sambalpuri POS taggers. The whole corpus size comprises of five domains, viz. literature, sports, tourism, entertainment, and miscellaneous. The highest corpus size is registered in the domain of entertainment i.e., approximately 40 k while the 'miscellaneous' section accounts for the lowest number of data.

4.2 Corpus Annotation

The whole Sambalpuri corpus[4] is annotated using the ILCI Ann Tool[5] [21] following the BIS-ILCI tagset (see Table 2) devised for Odia language since there is no tagset

[4] This is the very first POS tagset developed for Sambalpuri.

[5] http://sanskrit.jnu.ac.in/ilciann/index.jsp.

Table 1. Domain-wise corpus distribution

Domains	Tokens
Literature	30, 344
Sports	21, 121
Tourism	26, 767
Entertainment	40, 554
Miscellaneous	2, 424
Total	1, 21, 210

available for it. The BIS tagset is a hierarchical set designed by the POS Standardization Committee appointed by the Department of Information and Technology, Government of India. It has a total number of 11 categorical labels at the top level and 39 fine-grained labels for the annotation. The tagset is framed keeping in view both the fineness and coarseness or flat and hierarchical structures in view. The table below contains the nomenclatures of all the categories in the second column, annotation labels in the third and categorical IPA examples in the fourth.

Table 2. BIS parts of speech tagset adapted for Sambalpuri

Sl. No.	POS Categories	Annotation Labels	Examples of Sambalpuri in IPA
1	Noun	N	
1.1	Common	N_NN	pəʈər, pɪʈəl, bʰɑbnɑ, mʊnʊs
1.2	Proper	N_NNP	rɑm, hɪmɑləj, gəŋɑɖʰər meher bɪsʊəbɪɖjɑləj, səmbəlpʊr etc.
1.3	Verbal	N_NNV	pəʈʰɑ, pəhərɑ, ɖɛgɑ, nɑcbɑrʈɑ
1.4	Spatial & temporal	N_NST	ɑgke, pəcʰɑʈe, pəre, pʊrʊb, etc.
2	Pronoun	PR	
2.1	Personal	PR_PRP	mʊĩ, tʊɪ, ɑpən, se etc.
2.2	Reflexive	PR_PRF	nɪɟe etc.
2.3	Relative	PR_PRL	ɟɑhɑr, ɟɑhɑ̃kər, ɟenmɑnkər
2.4	Reciprocal	PR_PRC	nɪɟər bʰɪʈre, ɖʊhe etc.
2.5	Wh-word	PR_PRQ	kɪe, kɑhɑr, ken mɑne, etc.
2.6	Indefinite	PR_PRI	ənjər, kɛnsɪ, kehɪ etc.
3	Demonstrative	DM	
3.1	Deictic	DM_DMD	ɪ, se, ɪguʈɑkə, seguʈɑkə etc.
3.2	Relative	DM_DMR	ɟenguʈɑkə, ɟɑhɑr
3.3	Wh-word	DM_DMQ	kɑɳɑ, kenər, kenguʈɑkər etc.
3.4	Indefinite	DM_DMI	ʊnɪɑ̃, kɛnsɪ etc.
4	Verb	V	
4.1	Main	V_VM	sʊ, ɖəʊʈ, ɖɛkʰ etc.
4.1.1	Main	V_VM	

(*continued*)

Table 2. (*continued*)

4.1.2	Non-finite	V_VNF	kʰɑɪ kərɪ, nɑcɪ nɑcɪ, etc.
4.1.3	Infinitive	V_VINF	kʰɑɛbɑrke, kʰɑɪbɑr lɑgɪ, nɑcbɑr etc.
4.1.5	Gerund	V_VNG	kʰɑɪʈʰɪbɑr, kʰɑuʈʰɪbɑr etc.
4.2	Auxiliary	V_VAUX	ʊcɪʈ, ɖərkɑr, kərɪ, ʈʰɪbɑr etc.
5	Adjective	JJ	bʰəl, uʈʈəm, sũɖər etc.
6	Adverb	RB	
7	Postposition	PSP	sɑɲe, lekʰe, lɑgɪ etc.
8	Conjunction	CC	
8.1	Coordinator	CC_CCD	kã̃ hɛlɑɲe kɪ, kɑrən, ɑʊ, etc.
8.2	Subordinator	CC-CCS	ɟəɖɪ ʈebe, ɟeʈebɛlɛ seʈebɛlɛ, ɟe, bəlɪ etc.
8.3	Quotative	CC_CCS_UT	ɑɑre, həɛ lɒ, həgɒ, ɑgjã̃ etc.
9	Particles	RP	
9.1	Default	RP_RPD	bʰɪ, hĩ, ʈɔ etc.
9.2	Classifier	RP_CL	gʊʈe, ɖʊɪʈɑ, kʰə̃ɖe etc.
9.3	Interjection	RP_INJ	ʊɑh, həɛ, ɑh, oho etc.
9.4	Intensifier	RP_INTF	əʈɪ, kʰʊb, bəhʊʈ, ɟəbər etc.
9.5	Negation	RP_NEG	nɑĩ, nʊhe, nɪ, nɪhɑ etc.
10	Quantifiers	QT	
10.1	General	QT_QTF	ʈʰʊɖe, bɛsɪ, ʈɪke, gʊɖɑɖʊ etc.
10.2	Cardinal	QT_QTC	ek, ɖɒ, ʈɪn, cɑr etc.
10.3	Ordinal	QT_QTO	pəhɛlɑ, ɖʊsrɑ, ʈɪsrɑ etc.
11	Residuals	RD	
11.1	Foreign words	RD_RDF	languages of the other scripts except Odia
11.2	Symbol	RD_SYM	mathematical and other symbols (#, [, {, %, $, <, >, (,), *, @,)
11.3	Punctuation	RD_PUNC	(, ; : ‘’ “” :- etc.)
11.4	Unknown	RD_UNK	Tags that are left undecided
11.5	Echo word	RD_ECH	bɑgʰ-pʰɑg, kəʈɑ-cʰəʈɑ etc.

4.3 Data Size for the Taggers

The tabulated data (see Table 3) represents the different data sets applied to develop the statistical taggers. The total number of training data used for developing the taggers amounts to around 80 k. Initially, the tagger is trained with around 50 k with manually annotated data and later, the development set consists of 30 k which was automatically tagged and manually validated. After the training period, the testing is conducted with a set of approximately 13 k corpus size tokens.

Table 3. Training and testing data sets

Data sets	Tokens
Training	80, 288
Testing	12, 791

4.4 Developing POS Taggers

Two statistical taggers are developed for Sambalpuri; the first one is trained with SVM [22, 23] and the second is with CRF++ [24]. So far as the former is concerned, learning phase contains medium verbose (-V 2) and the mode of learning and tagging is set to left-right-left (LRL). The rest of the features like sliding window, feature set, feature filtering, model compression, C parameter tuning, Dictionary repairing and so on are set to the default mode. On the other hand, the latter is trained with the unigram method.

5 Issues and Challenges

This section is divided into three major sub-sections: corpus-related, human annotation-related and tagger-related issues.

5.1 Corpus-Related Issues

The issues pertaining to the corpus collection are vividly discussed: corpora collection, unavailability of Unicode encoding, non-standard usage of the language, different writing conventions and Hindi-like constructions.

Corpus Collection: A number of corpora have been developed for various languages like English and some European languages. Considering the situations in non-scheduled (lesser-known) Indian languages, it is quite unfavorable in comparison to the scheduled languages since some of the Indian institutions have either worked on or are presently developing language resources and technologies for the latter languages only. Because of the indifference of the government towards the lesser-known languages, the former are getting disempowered gradually. The institutions and projects that have worked for the corpus collection in scheduled Indian languages are IIIT-Hyderabad, CIIL-Mysore, ILCI-JNU and TDIL.

Unavailability of Unicode Encoding: Since low-density languages are less-resourceful or with no resource the software available for them are also less in number. This leads to the non-Unicode encodings which is not favorable for the development of NLP applications. The whole corpus has been converted into UTF-8 encodings using Akruti Text Converter[6]. There are different linguistic issues in the corpus itself such as non-standard usage, non-uniform Orthographic forms and Hindi-like constructions.

[6] https://22bc339da9ca3e2462414546a715752e4c2c5e0d.googledrive.com/host/0B5rBGd680WZFemVLa3RxY0preE0/AkrutiUnicode.

Non-standard Usage: Sambalpuri is not a scheduled Indian language and is written and spoken with varying standards in different regions of the western and south-western Odisha. For example: Sambalpuri, Bargadia (spoken in Bargarh), Bolangiri/a (spoken in Bolangir district), Sundargadi/ia (spoken in Sundargarh), Deogarhia (spoken in Deogarh region), Boudia (spoken in Bouddha district) etc. There are some dialectical variations among the people of Sambalpuri speaking track. The table (see Table 4) demonstrates dialectal variations of Sambalpuri with reference to negative morpheme 'no', adverbs 'now' and 'this way'. Lexical similarity within the varieties of Sambalpuri is considerably high which ranges from 90 to 95% [5]. This similarity matrix was made by comparing Bargarhi, Bolangiri and Jharsuguda varieties with Sambalpuri.

Table 4. Dialectal variations in Sambalpuri (adapted from [25])

Variety of Sambalpuri	Negative 'no'	Adverb 'now'	Adverb 'this way'
Bargarh	nʊhe/nɪhe	ɪhaɖ̪e/ɛcʰɛn	ɪaɖe/ɪpʰɑle
Bolangir	nĩ	ɛkʰɛn	
Kalahandi	nĩ	ɛkʰɛn	ɪbaʈe
Sambalpur	nɪhe/nʊhe	ɪcʰnɪ	ɪaɖe
Sundargarh	nĩ		
Bouda		ɪgəɖɪ	ɪaɖkʊ

Different Orthographic Conventions: A large number of words in Sambalpuri has different orthographic conventions; especially the ligatures. In Sambalpuri, there are several writing conventions used for a given word form because of the non-uniform usage of language.

For instance, in the following examples two forms are used for one word with two of them having different POS labels with the change of form.

କାଞ୍ଚି N_NN, କାଁଜେ DM_DMQ

କନ୍‌ଡକ୍‌ଟର N_NN, କଣ୍ଡକ୍ଟର୍ N_NN

କାନ୍‌ଚନ N_NN କାଞ୍ଚନ N_NNP

This non-standard usage of the words creates issues during both manual and automatic annotation since their POS labels vary with the varying conventions.

Hindi-like Constructions: Sambalpuri is more like Hindi than Odia which accounts for the fact that the western region, where it is spoken, is situated just adjacent to Chattisgarh and Jharkhand where the influence of Hindi is largely felt. In the examples instantiated below /bɑʊəɟʊɖ̪/ and /ke/ are postpositions as used in Hindi while the Hindi-like indefinite and reflexive pronouns are also used.

For instance,

/bɑʊəʃʊd̪/ PSP
/hər ek/ PR_PRI /ke/ PSP
/əpnɑ/ PR_PRF /əpnɑr/ PR_PRF

5.2 Issues Pertaining to Human Annotation

One of the prominent challenges is that which pertains to the annotation of the corpus. For annotation of a voluminous corpus and to maintain consistency, one needs to have a standard tagset. Owing to the fact that a large number of languages like Sambalpuri being less-described or less-studied, it is quite daunting to devise a tagset. If one adopts and adheres to the tagset devised for a language of close proximity, then they may either compromise with the saliency of the linguistic data or may end up filling different slots for labels and not researching by delving deep into some interesting structures. For instance, there are large numbers of homophonous words that can neither be included in the reduplicated nor can they be labelled as echo.

Reduplicated Expressions: Generally, in Indian languages the reduplicated expressions follow the meaningful word. Contrastingly, in Sambalpuri many of the reduplicated parts precede the meaningful words (see Sect. 3.3). For instance, in the conjunct verb (adjective + finite verb), /cʰɪcʰɪ/ is the meaningless reduplicated part which is preceding the meaningful part /bɪcʰɪ/ 'scattered'.

For example,

cʰɪcʰɪ\RD_ECH bɪcʰɪ\JJ heɪcʰən 'have got scattered'

Similarly, in the following verbal reduplication, the meaningless part is preceding the verbal part.

For example,

kʊt̪\RD_ECH kʊt̪eɪ\V_VM d̪ɛlɑ\V_VM_VF 'has tickled'

These kinds of constructions pose significant linguistic challenges for the human annotators as to how to label them and so is for the statistical tagger.

Verb-less Constructions: In Sambalpuri and many sister languages such as Odia, Bengali, Assamese [26] verb-less constructions or covertly present verbs are commonly used. These constructions are used with adjectives in place of verbs. Therefore, the tagger also labels some of these adjectives as finite verbs because of the annotation of these constructions in the training data.

For example,

sɑsʊt̪\N_NNP bɑbʊr\N_NN cɑnvɑs\N_NN ʊsɑr\JJ pɪsɑr\RD_ECH .\RD_PUNC
"Saswat Babu's canvass is quite large."

In the above example, /ʊsɑr\JJ pɪsɑr\ RD_ECH/ is the reduplicated adjectival phrase which satisfies the need of the verb.

Onomatopoeic Constructions: Onomatopoeic words are the imitation of a sound associated phonetically with its describing referent. These following expressions are

parts of the multi words because individually these words do not have meaning, but when combined they are manner adverbs. As per the ILCI guideline, if we annotate the first sound as noun and the following words as echo-words (RD_ECH), we are missing relevant linguistic information.

For instance,

bʰɛ̃ bʰɛ̃ 'loudly'
ɟʰɒ ɟʰɒ 'heavily'
d̪ʰɔ̃ pɔ̃ 'gasping'
bʰɒ bʰɒ 'bark'

Agglutination of Classifiers with Postpositions: Agglutination (see Sect. 3.2) is one of the common features in Odia [10] and Sambalpuri along with some IA languages like Bengali and Marathi [18]. In Sambalpuri, one of the peculiar constructions with agglutination is that the classifiers and postpositions agglutinate with each other which is also rarely found in the most agglutinating Dravidian languages. Here, to annotate these constructions as classifiers (RP_CL) or postpositions (PSP) is quite difficult.

For instance,

/bɑgɪr-ʈɑ/ 'as-CL'
/lekʰeʈɑ/ 'like-CL'

Similarly, in the example below, it is quite difficult to decide the annotation labels for both the human and automatic annotation. The reason is the complexity in deciding the head label of the words. The word /d̪ʊɪ-ʈɑ/ comprises of two components, a cardinal and a classifier morpheme. Both of these categories have separate labels in the BIS scheme for Sambalpuri. Therefore, if one annotates the word as cardinal, they are compromising with the other label or linguistic information.

For instance,

d̪ʊɪ-ʈɑ (d̪ʊɪ\QT_QTC ʈɑ\RP_CL)

5.3 Issues Related to Automatic Annotation

These issues are mostly pertained to tagger-related ambiguities and some other linguistic errors.

Ambiguity Issues: The data (see Table 5) represented below demonstrates that there are different types of ambiguous sets of classes and their accuracy rates. All the ambiguity classes are divided into 244 classes and they are generated automatically by the SVM tool.

Two-label Sets: This section includes the ambiguous words with two conflicting labels. The most commonly ambiguous tags are coordinating-subordinating conjunctions, coordinating conjunction-general quantifier, deictic-interrogative demonstrative and so on.

ɑʊ (CC_CCD or QT_QTF)

Table 5. Ambiguity Classes

Classes	Label Sets
2 Sets:	CC_CCD_CC_CCS, CC_CCD_QT_QTF, DM_DMD_DM_DMQ
3 Sets:	JJ_N_NST_V_VM_VF, RD_ECH_V_VM_V_VM_VNF, RP_NEG_V_VM_V_VM_VF
3 Sets>	V_VAUX_V_VM_V_VM_VF_V_VM_VNF, RP_INJ_RP_NEG_V_VM_V_VM_VNF, RD_UNK_RP_INJ_RP_RPD_V_VM_V_VM_VNF

For example,

mʊĩ ɑʊ\CC_CCD mɒr bɑpɑ "I and my father"
mə̰ʈe ɑʊ\QT_QTF kʰɑnɑ ɖə̰rkɑr "I need some more food".

In the above examples, the first one suggests that the word /ɑʊ/ is a coordinating conjunction coordinating two noun phrases while the second one states that it is a general quantifier used as a pre-modifier of the following head noun.

Three-Label Sets: This section contains the ambiguous words having three labels. The most commonly ambiguous tags are most-expectedly adjective-temporal nouns-finite, negative-main-finite verb and so on.

bɑhɑr (V_VM or N_NST or JJ)

For instance,

bɑhɑr\V_VM_VF gʰərʊ\N_NN "Come out of the house".
se pəlalɑ bɑhɑr\JJ gʰərʊ\N_NN "He went away from the front room".
bɑhɑr-ke\N_NST ɑs\V_VM_VF "Come to outside".

In the instances mentioned above, the word form /bɑhɑr/ has three different POS labels. The first one is annotated as a finite main verb as the sentence is an imperative sentence and the covert subject is the second person pronominal. In the second example, it is labelled as an adjective as it modifies the following noun whereas the third one is a spatial noun as it refers to a location.

More than Three-Label Sets: The words having more than three labels are encapsulated in this part. For instance, main-auxiliary-nonfinite-finite verbs, unknown-interjection-default particle and so on.

kərɪ (V_VAUX or V_VM or V_VM_VF or V_VM_VNF)

For instance,

kʰɑɪ\V_VM kərɪ\V_VAUX ɑsle\V_VM_VF
kɑm\N_NN kərɪ\V_VM ɑsle\V_VM_VF
kərɪʈ̰ʰɪlɑ\V_VM_VF
kʰɑɪkərɪ\V_VM_VNF ɑsle\V_VM_VF

The verbal word form /kərɪ/ has more than three labels in the corpus and which is rightly so. It can be used as main, auxiliary, finite and non-finite verbs as instantiated in the above examples.

6 Results and Discussion

In spite of the different issues and challenges, statistical taggers developed for Odia achieves accuracies of 94% and 89% by SVM and CRF++ respectively (Behera [10]). The results (SVM 83% and CRF++ 71.56%) for Sambalpuri are comparatively lesser than Odia which accounts for the fact that Sambalpuri has no standardized orthographic convention and hence different regional varieties use the language in their own ways.

The first and foremost point to emphasize for a low-density language is the large-scale writing on the social media using its own script with a soul objective of developing language resources. If there is less availability of the corpus, then one can also take the assistance of mathematical modelling to achieve a higher accuracy rate. So far as the tagset-related issues are concerned, one can take the labels already used by a closely-related language spoken in the region for annotation job by incorporating it on their convenience. With regard to issues in annotation, one can take the exemplary labels from different tagsets developed for Indian languages. For example, we can incorporate WRB label from the IIIT-Hyderabad for the interrogative adverb. The reduplicative expressions need to be considered seriously as they are the most vital parts of the language and they behave quite differently in Sambalpuri. Therefore, it can be averred that labels for reduplication (RD_REDP), possessive pronouns (PR_POS) and demonstratives (DM_POS), interrogative adverbs (WRB) can be introduced. For handling agglutination, a stemmer or a lemmatizer could be used with statistical POS taggers. For punctuations, fine-grained labels should be incorporated based on their functions in a given context as they can be used as coordinators, section headers, list item markers and so on. Finally, the standardization of the language would help solve many of the issues by providing consistency in both the human and statistical annotations.

7 Conclusion

In this paper, we have discussed about different issues and challenges in terms of both corpus collection, annotation and tagger-related issues in detail for a less-resourced language, i.e. Sambalpuri. The results (SVM 83% and CRF++ 71.56%) of the statistical taggers for Sambalpuri in the present study would not only prove to be beneficial for its own future NLP research and development but also would be advantageous for any other morphologically-rich less-resourced language from around the world. At a later stage, we can further use a lemmatizer or stemmer to handle the agglutination issue and incorporate some of the solutions proposed in the research. Furthermore, these POS taggers could be potentially used for developing morph analyzer, chunker, parser and hopefully for enhancing the accuracy of machine translation. For its future development, a full-fledged online lexical dictionary using Language Explorer, Lexique Pro & Toolbox can also be prepared.

References

1. McEnery, T., Baker, P., Burnard, L.: Corpus resources and minority language engineering. In: LREC (2000)
2. Ostler, N.: Language technology and the smaller language. ELRA Newsl. **4**(2) (1999)
3. Abbi, A.: A Manual of Linguistic Fieldwork and Structures of Indian Languages, vol. 17. Lincom Europa (2001)
4. Kushal, G.: Case and agreement in Sambalpuri. M. Phil. Thesis, Centre for Linguistics, Jawaharlal Nehru University, New Delhi, Delhi (2015)
5. Mathai, E.K., Kelsall, J.: Sambalpuri of Orissa, India: A Brief Sociolinguistic Survey. SIL International (2013)
6. Tripathy, B.: Sambalpuri semantics. Graduate Thesis, Sambalpur University, Sambalpur (1984)
7. Behera, P. Dash, B.N.: Documenting Sambalpuri-Kosli: the case of a less-resourced language. Indian J. Appl. Linguist. (IJOAL). Bahri Publications (0379-0037), June 2017. (accepted)
8. Padhy, H.H., Mohanty, S.: Designing hybrid approach spell checker for Oriya. Int. J. Latest Trends Eng. Technol. **2**(4), 156–160 (2013)
9. Jena, I., Chaudhury, S., Chaudhry, H., Sharma, Dipti M.: Developing Oriya morphological analyzer using Lt-Toolbox. In: Singh, C., Singh Lehal, G., Sengupta, J., Sharma, D.V., Goyal, V. (eds.) ICISIL 2011. CCIS, vol. 139, pp. 124–129. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19403-0_20
10. Behera, P.: Odia parts of speech tagging corpora: suitability of statistical models. M. Phil. Thesis, Centre for Linguistics, Jawaharlal Nehru University, Delhi (2015)
11. Behera, P.: Evaluation of SVM-based automatic parts of speech tagger for Odia. In: Proceedings of WILDRE-3 (LREC-2016), Portoroz, Slovenia, pp. 32–38 (2016). ISBN: 978-2-9517408-8-4
12. Behera, P.: An experimentation with the CRF++ parts of speech tagger for Odia. Lang. India **17**(1) (2017). ISSN: 1930-2940
13. Ojha, A.K., Behera, P., Singh, S., Jha, G.N.: Training & evaluation of POS taggers in Indo-Aryan languages: a case of Hindi, Odia and Bhojpuri. In: Proceedings of LTC-2015, Poland, pp. 524–529 (2015)
14. Behera, P.: Issues and challenges in corpus collection and annotation of Sambalpuri: the case of a lesser-known language. Language Forum, Bahri Publications, June 2018. ISSN 0253-9071. (accepted)
15. Bhattacharya, T.: The structure of the Bangla DP. Doctoral Dissertation, University College, London (1999)
16. Neukom, L., Patnaik, M.: A Grammar of Oriya. Seminar für Allgemeine Sprachwissenschaft der University, Zürich (2003)
17. Shukla, S.: Bhojpuri Grammar. Georgetown University Press, Washington, D.C. (1981)
18. Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharyya, P., Jha, G.N.: A common parts-of-speech tagset framework for Indian languages. In: LREC (2008)
19. Abbi, A.: Reduplication in South Asian Languages: An Areal, Typological and Historical Study. Allied Publishers Pvt. Ltd., Chennai (1992)
20. Jha, G.N., Hellan, L., Beermann, D., Singh, S., Behera, P., Banerjee, E. Indian languages on the TypeCraft platform - the case of Hindi and Odia. In: LREC, Iceland (2014)

21. Kumar, R., Kaushik, S., Nainwani, P., Banerjee, E., Hadke, S., Jha, G.N.: Using the ILCI annotation tool for POS annotation: a case of Hindi. In: 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012), New Delhi, India, March 2012
22. Joachims, T.: Making large scale SVM learning practical. Universität Dortmund (1999)
23. Giménez, J., Màrquez, L.: Technical Manual v1.3. Universitat Politecnica de Catalunya, Barcelona (2006)
24. Kudo, T.: CRF ++: Yet Another CRF Toolkit (2013). http://crfpp.sourceforge.net/ptojrcts/crfpp/. Accessed 10 July 2015
25. Patel, K.: A Sambalpuri Phonetic Reader. Menaka Prakashani, Sambalpur (undated)
26. Masica, C.P.: The Indo-Aryan Languages. Cambridge University Press, Cambridge (1993)