# Towards a Part-of-Speech Tagger for Awadhi : Corpus and Experiments

**Abdul Basit, Ritesh Kumar**
Department of Linguistics
Dr. Bhim Rao Ambedkar University, Agra
basitansari03@yahoo.com, riteshkrjnu@gmail.com

**Abstract**
Awadhi is an Indo-Aryan language, spoken in the eastern region of Uttar Pradesh by approximately 38 million native speakers. However, despite this large number of speakers, it is highly lacking in language resources like corpus, language technology tools, guidelines etc till date. This paper presents the first attempt towards developing an annotated corpora and a POS tagger of the language, The corpus is currently annotated with part-of-speech tags. Since there is no earlier tagset available for Awadhi, the POS tagset for the language was developed as part of this research. The tagset is a subset of the BIS scheme, which is the national standard for the development of POS tagsets for Indian languages.

**Keywords:** Awadh, POS Annotation, BIS, Corpus development, Less-resourced language

## 1. Introduction

Awadhi is an Indo-Aryan language, spoken in the eastern region of Uttar Pradesh viz. Lucknow, Raebareli, Sitapur, Unnao, Allahabad, Faizabad, Sultanpur, Behraich and Pratapgarh etc. According to 2001 census, there are 38 millions native speakers of Awadhi language. It is the official language of Nepal and Fiji. Awadhi writing system follows Devanagri, Kaithi and Perso-Arabic script.

In present scenario of India, there are several attempts to collect the corpus of Indian languages and few corpora are available in some of the major languages of India. However, there is no corpus available for Awadhi till now. In the present research, the data of Awadhi language is collected from Eastern region of Uttar Pradesh. In this research, I have developed a corpus with approximately 70,000 tokens. Approximately 20,000 tokens of the corpus data has been annotated with the POS information. It is the first POS-Tagged corpus of Awadhi language. I have also developed the first POS tagset of Awadhi based on the general BIS tagset for Indian language.

The coherent ratio through different varieties of the langage is very rich, but other elements such as affixes, auxiliaries, address terms and domian specific terms differ a lot. The word order of Awadhi is Subject Object Verb (SOV). The use of postposition like मा, से , का etc. indicate possession in Awadhi. Final noun head, two genders ; Musculine and Feminine, clause constituents indicated by case marking. The verbal affixation marks person, number and gender of subject and object. There is an ergative less non-tonal language. There are 30 consonants and 8 vowels phonemes in Awadhi. The writing system fallows Devanagri and Perso-Arabic ascript. The morphological typology of Awadhi language is fusional. (Awadhi/Ethnolgue)

## 2. Development of POS-Tagged Corpus

In this section, we discuss the process of the development of the post-tagged corpus of Awadhi. It includes the methods of data collection, sources of data, format of data and metadata for current research. It also discusses the challenges and issues in Optical Character Recognition (OCR) of Awadhi texts using a Hindi OCR system and how we worked around the problems. We also discuss the part-of-speech (POS) tagset of Indian languages approved by the Bureau of Indian Standards (BIS) and the POS Annotation tool that we have used for annotating the data.

### 2.1. Corpus Collection

Corpus provides an empirical base for various linguistic observations, hence, it is a primary source of data for the purpose of linguistic studies and for developing various tools for Computational Linguistics and Natural Language Processing. The ideal aim of data collection is to include as much diversity of a language as possible. As such it is carried out to include millions of words collected from different domains. The current research, however, aims to collect at least hundred thousand tokens of Awadhi language for the corpus formation.

### 2.2. Source of Data

The data for the current research has been collected from Uttar Pradesh Hindi Sansthan's Library and various publication house in Lucknow. The corpus data has been collected from primary sources i.e., textbooks, short stories and novels. Some of the sources which has been used for data collection include

- Chandawati
- Nadiya Jari Koyla Bhai
- Tulsi Nirkhen Raghuvar Dhama

The current corpus includes data from these novels published in Awadhi.

**Lack of Resource –** Despite being spoken by a large population, Awadhi lacks electronic as well as other kinds of resources. There is only one website named as *Awadhi kay Arghan* (www.awadhi.org.), where, a very limited number of short stories and poetries are available in Awadhi. Some Facebook pages like *Awadhibhasha, Awadhi Wikipedia etc* claim to promote the cause of Awadhi but they hardly contain writings in Awadhi. There are not any regular electronic newspapers, blogs and magazine available in Awadhi language. The language also does not any published grammar or dictionary available. As such it is a rather challenging task to collect data for the language and even a minimal corpus could prove be very useful.

## 2.3. Method of corpus creation

In order to expedite the process of creating data, we did not digitise the texts manually. Instead a pipeline of scanning, OCR and proofreading was followed. We expected this process to be much quicker than manually typing out the contents to digitise them. However, this method also had its own set of challenges, which is discussed in the next section.

## 2.4. Challenges in Optical Character Recognition (OCR) of Awadhi Texts

The data collection process for current research was quite challenging from several perspectives. The very first and basic challenge was the absolute lack of corpus of Awadhi. And so it naturally follows that OCR system is not being developed for Awadhi. As a result, a lot of spelling errors were found in the OCRed Awadhi corpus data when Awadhi texts were scanned using Devanagri OCR system. Some of the most common spelling errors in OCR are mentioned in Table 1.

| Spelling Error | Correct words/Matras |
|---|---|
| ो (तो) | ौ(तौ |
| े(ले) | ै(लै) |
| द। द 7 | दादा |
| द् याखौ | द्याखौ |

Table 1 : Most common spelling

As we could see, the errors seem to be largely because of the absence of such words in Hindi and the presence of a very closely-related but quite different word in Awadhi it could be hypothesised that these errors might be because of auto-correction by the 'Hindi' OCR system. Such errors necessitated a manual proofreading of the corpus. The proofreading was carried out a Java/JSP-based in-house editing tool 'editit'.

## 2.5. The Corpus Editing tool: Editit

A corpus editing tool, Editit, was developed using Java/JSP at the backend and runs on Apache Tomcat 8.5 web server. This tool helped in proofreading and correcting the errors that creeped into the corpus data after OCR.

## 2.6. POS tagsets for Indian Languages

The Penn Tree bank tagset has emerged as for POS tagging of western languages. But Indian languages is much more morphologically rich features. There are a number of POS tagsets designed by several research groups working on Indian Languages. These are, IIT (ILMT) tagset, LDC-IL tagset (Chandra, kumawat & Srivastava, 2014), AUKBC tagset, JNU Sanskrit tagset (JPOS) (Gopal, 2009), MSRI tagset (Baskaran et al., 2008), CIIL- Mysore tagset and BIS tagset (Chaudhary, 2010) is one of them.

Leech and Wilson (1999) espoused the case of standardization of tagset for their reusability of anointed corpora and interoperability across different languages. The result of their effort was EAGLES guidelines. To Achieve the same results BIS has been adopted as the standard for Indian Languages.

BIS is a national-level body that decides on the standard and since this framework (from which the Awadhi tagset was derived) has been approved by BIS, it is now a national standard and is expected to be used by anyone working on POS tagging of Indian languages – and this is the main motivation of using this tagset. Moreover, the BIS framework allows to derive tagsets for different Indian languages; however, the other tagsets are neither accepted as national standards nor are they developed as generic framework, BIS framework was used for building Awadhi tagset.

## 2.7. Annotation of the data: BIS Tagset

The Bureau of Indian Standards (BIS) Tagset has recommended the use of a common tagset for the part of speech annotation of Indian languages. The tagset, incorporating the advice of the experts and the stakeholders in the area of natural language processing and language technology of Indian languages, has to be followed in the annotation tasks taking place in Indian languages (Chaudhary and Jha, 2011).

Since there is no earlier tagset available for Awadhi, a POS tagset for the language was developed as part of this research. The tagset is a subset of the general BIS tagset. It is used for the POS tagging of Awadhi corpus of approximately 20 thousand tokens. The tagset has 32 different categories including punctuation, residual and unknown category. The complete tagset is given in Table 2.

| S.NO. | Categories | Subtypes Level 1 | Annotation Convention | Exam-ples |
|---|---|---|---|---|
| **1** | **Noun** | **N** | **N** | **मेहरारु ,किताब दारोगा ,मनसेदु** |
| **1.1** | | Common | N_NN | चश्मा, गिलास, बासन, डाक्टर |

| | | | | |
|---|---|---|---|---|
| **1.2** | | Proper | N_NNP | अब्दुल, योगेश, रीना, अनम |
| **1.3** | | Nloc | N_NST | ऊपरै, नीचे, आगै, पीछै |
| | | | | |
| **2** | **Pronoun** | **PR** | **PR** | **वुइ, तुमे, यहे** |
| **2.1** | | Personal | PR_PRP | वुइ, तुमरे, |
| **2.2** | | Reflexive | PR_PRF | अपन, हमरे, खुदै |
| **2.3** | | Relative | PR_PRL | जौ, जिस, जबै |
| **2.4** | | Recipro-cal | PR_PRC | दुनौ, आपसै |
| **2.5** | | Wh-word | PR_PRQ | कबहूँ, काहे, का |
| **2.6** | | Indefinite | PR_PRI | केउ, किस |
| | | | | |
| **3** | **Demon-strative** | | **DM** | **हिंया, हुंआ, जौ** |
| **3.1** | | Deictic | DM_DMD | हिंया, हुंआ |
| **3.2** | | Relative | DM_DMR | जे, जौन |
| **3.3** | | Wh-word | DM_DMQ | के, काहे |
| **3.4** | | Indefinite | DM_DMI | काउनौ, किस |
| | | | | |
| **4** | **Verb** | | **V** | **गवा, रहन** |
| **4.1** | | Main | V_VM | कीन, कै, गवा |
| **4.2** | | Auxiliary | V_VAUX | रहन, रहै, होय, है |
| | | | | |
| **5** | **Adjective** | | **JJ** | **बड़ा, अच्छै** |
| **6** | **Adverb** | | **RB** | **तेजी,** |
| **7** | **Postposi-tion** | | **PSP** | **मा, से, का** |
| **8** | **Conjunc-tion** | | **CC** | **औ, अउर, बल्कि** |
| **8.1** | | Co-ordi-nator | CC_CCD | औ, बल्कि |
| **8.2** | | Subordi-nator | CC_CCS | तौ, कि |
| | | | | |
| **9** | **Particles** | | **RP** | **बहुत, हे, ना, भी** |
| **9.1** | | Default | RP_RPD | भी, ही |
| **9.3** | | Interjec-tion | RP_INJ | अरै, हे, वाह |
| **9.4** | | Intensi-fier | RP_INTF | बहुत |
| **9.5** | | Negation | RP_NEG | नाही, ना, बिना |

| | | | | |
|---|---|---|---|---|
| **10** | **Quanti-fier** | | **QT** | **तनिक, एक, पहिला** |
| **10.1** | | General | QTF | तनिक, बहुतै, कुछै |
| **10.2** | | Cardinals | QT_QTC | एक, दुई, छ |
| **10.3** | | Ordinals | QT_QTO | पहिला, दुसर का, तीसर का |
| | | | | |
| **11** | **Residuals** | | **RD** | |
| **11.1** | | Foreign word | RD_RDF | **Other than script of the origi-nal text** |
| **11.2** | | Symbol | RD_SYM | **$,&,*, (,)** |
| **11.3** | | Punctua-tion | RD_PINC | **., :, ;, “,’, । , ?,!,** |
| **11.4** | | Unknown | RD_UNK | |
| **11.5** | | Echo-words | RD_ECH | **खाना-वाना, कुर्सी-उर्सी** |

Table 2 : POS Annotation Scheme of Awadhi

As one would notice, BIS tagset bears close resemblance to the LDC-IL tagset. In addition to one type of a category. It also introduces another subtype. BIS tagset groups together unknown, punctuation and residual in one top-level tag – Residual while LDC-IL tagset had three different tags for these. Noun and Pronoun in the two tagsets are almost identical in the two tagsets. Verb (V), too, has the same subtypes – main verb (VM) and auxiliary verb (VAUX). Adjective and Adverb has no subtype whereas we have two new categories in BIS tagset – one is conjunction (CC) which has two subtypes namely coordinator (CCD) and subordinator (CCS). These subtypes were grouped under particle (RP) in LDC-IL tagset. As a result, Particles (RP) in BIS contains Default(RPD), Classifier(CL), Interjection(INJ), Intensifier(INF) and Negation(NEG) as its subtypes. The other category not in BIS tagset is numerals(NUM) – it is replaced by Quantifier(QT), with General(QTF), Cardinal(QTC) and Ordinal(QTO) as its subtypes. Expect for the three categories of adjective, adverb and postposition , all the categories have two or more sub-categories. Moreover, the category of residual, although not part of the language, it is part of the text which is to be annotated and so included in the tagset. See (Appendix)

### 2.8. Corpus statistics

Overall, the corpus currently consists out 8,532 sentences, amounting to a total of 95,717 tokens. Out of these, 21,256 tokens are currently tagged with part-of-speech information. We are actively working on the development of this corpus and we hope to get 100k pos-tagged token over a period of next few months.

## 3. Automatic POS Tagger : Experiments and Results

In order to develop an automatic part-of-speech tagger for Awadhi, we have experimented with a tagged corpus of 21,526 tokens that was tagged by a single annotator using the tagset discussed above.

We experimented with 2 classifiers – Decision Trees and Support Vector Machines (SVM) – using the following set of features -

**Word-level features** : We used the current word, previous 2 words and next 2 words as features

**Tag-level features** : We used the tags of previous 2 words as features.

**Character-level features** : We use the first three characters (prefixes) and last three character (suffixes) as features for training

**Boolean features** : In addition to the above features, we also used the following additional features – *has_hyphen* (1 if the word has hyphen in it), *is_first* / *is_second* (1 if the word is the first / second word in the sentence), *is_last* / *is_second_last* (1 if the word is the last / second last word in the sentence) and *is_numeric* (if the word is a number).

Using these features, the performance of the two classifiers are summed in Table 3 below

| **Classifier** | **Decision tree** | **SVM** |
|---|---|---|
| Precision | 0.75 | 0.78 |
| Recall | 0.75 | 0.78 |
| F1 | 0.75 | 0.78 |

Table 3 : Comparison of 2 POS taggers for Awadhi

As is pretty obvious, both the classifiers suffer from a lack of sufficient amount of data. And we expect the results to move closer to the current state-of-the-art in POS taggers as more data comes in.

## 4. Summing Up

In this paper, we have discussed the creation of a corpus of approximately 95k tokens in Awadhi and the POS-annotation of approximately 26k tokens. We have also discussed the development of an automatic POS tagger for the language which gives a best F1 score of 78 %. The low score could be explained by the minimal amount of data available for training the system – this is expected to improve as more data becomes available. This is a work in progress and over a period of next few months we hope to develop a bigger corpus as well as a better POS tagger.

## 5. Bibliographical References


AU-KBC tagset. AU-KBC POS tagset for Tamil. Retrieved from http://nrcfosshelpline.in/smedia/images/downloads/Tamil_Tagset-opensource.odt

Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharya, P., Choudhury, M., Jha, G.N., Rajendran, S., Sravanan, K., Sobha, L. and Subbarao K.V.S. (2008).Designing a common POS-Tagset Framework for Indian Languages. In Proceeding of VI workshop on Asian Language Resources, IIIT, Hyderabad.

Chaudhary, Narayan and Girish Nath Jha. (2011). Creating multilingual parallel corpora in Indian Languages. In Proceedings of the 5th Language and Technology Conference : Human Language Technology as a challenge for computer science and linguistics, pages 85 – 89, Poznan, Poland

Leech, Geoffrey and Wilson, Andrew. (1999). [Edited version of Eagles Recommendations for the Morphosyntactic Annotation of corpora. (1996): at http://www.ilc.cnr.it/EAGLES96//annotate/annotate.html.]

Nitish Chandra, Sudhakar Kumawat, Vinayak Srivastava (2014). Various tagsets for indian languages and their performance in part of speech tagging Proceedings of 5 th IRF International Conference, Chennai, 23rd March. 2014 http://www.digitalxplore.org/up_proc/pdf/55-139590032413-17.pdf

Jha, Nath, Girish., Madhav Gopal and Diwakar Mishra. (2009). Annotating Sanskrit Corpus: Adapting IL-POST. Springer Heidelberg Dordrecht, London New york.

IIT-Tagset, A parts-of-Speech tagset for Indian Languages. Retrieved from http://shiva.iiit.ac.in/SPASAL2007/iiit_tagset_guidelines.pdf(30-12-2017)

www.ethnologue.com/language/awa(27-12-2017)