INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Natural Language Processing Resources for the Kashmiri Language

Nawaz Ali Lone[1], **Kaiser J Giri**[2]*, **Rumaan Bashir**[2]

1 Research Scholar, Department of Computer Science, Islamic University of Science & Technology, Kashmir
2 Associate Professor, Department of Computer Science, Islamic University of Science & Technology, Kashmir

OPEN ACCESS

Received: 30-09-2022
Accepted: 11-10-2022
Published: 16-11-2022

Citation: Lone NA, Giri KJ, Bashir R (2022) Natural Language Processing Resources for the Kashmiri Language. Indian Journal of Science and Technology 15(43): 2275-2281. https://doi.org/10.17485/IJST/v15i43.1964

* **Corresponding author**.

kaiser.giri@islamicuniversity.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The main objective of this paper as a maiden attempt is to identify the basic resources necessary for undertaking Natural Language Processing (NLP) specific research activities pertaining to Kashmiri language. The paper also deliberates on key issues related to Natural Language Processing of Kashmiri language such as complex linguistic phenomena, the lack of standard linguistic tools, documented as well as standardized resources and the influence of some dominant languages mostly Urdu and English on Kashmiri language. **Methods:** As there is no substantial work reported in literature specific to NLP of Kashmiri language, a holistic research strategy was adopted to explore the possible sources as potential means for creation of basic resources to undertake the NLP research for Kashmiri language. **Findings:** After thorough investigation, it was observed that there has been some trivial work reported in the literature related to Machine Translation of Kashmiri language. Further there are few newspapers published in Kashmiri language which can be used as a means for creation of Kashmiri corpus. Moreover crowdsourcing could be used a potential means for development of digital linguistic resources for Kashmiri language. **Novelty:** The present study is a maiden attempt towards identification of NLP resources for Kashmiri language and will be of immense importance for the research community interested to work for development of Kashmiri language in digital domain.

Keywords: Natural Language Processing; Transliteration; Kashmiri Language; Scheduled Languages; crowdsource; Tag set; P-o-S Tagging

1 Introduction

The field of Natural Language Processing (NLP) is an emerging area of research and has thus received significant attention of researchers over past few decades[(1)]. It consists of diverse language specific research aspects such as Text and Speech Processing, Morphological Analysis, Syntactic Analysis, Lexical Semantics, Relational Semantics, etc. The field of Natural Language Processing has immense research potential owing to its applications across diverse domains such as Education, Healthcare, Military, and

Business etc. Therefore extensive NLP specific research work has been reported in the literature for most of International as well Indian languages[2].

The intervention of recent Artificial Intelligence and Machine Learning based practices furthermore have revolutionized almost every sphere of life and the field Natural Language Processing is no exception. These practices have resulted into more efficient solutions with improved performance. Since basic digital resources such as familiarity with linguistic structure, corpus etc. is a basic requirement for working of any cotemporary Artificial Intelligence and Machine Learning based NLP system, therefore their identification and availability is of significant importance.

Since there is no work reported in literature pertaining to NLP resources specific to Kashmiri language, the main objective of this research paper is to present a comprehensive survey of digitized language processing or NLP work related to the Kashmiri language, which has been declared as a scheduled language in the Eighth Schedule of the Indian Constitution[3]. The paper further highlights various NLP resources available for Kashmiri language besides deliberating on various relevant issues specific to Natural Language Processing of Kashmiri language. The work presented will serve as compiled information source for research community about research profile of the Kashmiri Language as has been done with other Indian Languages[4]. The paper is organized as follows:

Section 2 provides a brief about the Kashmiri Language.

Section 3 discusses NLP specific research resources available for the Kashmiri Language.

Section 4 provides the state of research on NLP for the Kashmiri Language.

Section 5 discusses issues concerning Kashmiri as a Language.

Section 6 provides the conclusion of the paper.

2 Brief about the Kashmiri Language

Kashmiri or Koshur[5] which belongs to Dardic subgroup of Indo-Aryan languages, inherits its label or name from the area where it is being used Figure 1. There are about 7 million Kashmiri speakers as per the census 2011, who primarily belong to Kashmir Valley, Chenab valley[6], Neelam valley, Leepa valley & district of Haveli[7,8]in the territory of Jammu and Kashmir. It is one among 22 Indian scheduled languages[9]. Although, the Perso-Arabic script is more common in use at present, it is primarily written in modified Perso-Devanagari scripts.

Fig 1. Descendance tree of Kashmiri Language

The Kashmiri language is one of the 22 scheduled Indian languages that has recently been declared as official language of Jammu & Kashmir[10]. For the development of the Kashmiri language, Jammu & Kashmir Government has made it as a compulsory subject in school education up to the secondary level[11] since November 2008. It is also one of the fifteen languages imprinted on Indian currency paper notes.

Kashmiri language is written using three orthographic systems, which include the Sharada Script, the Devanagari Script, and the Perso-Arabic Script. Quite often the Roman script is now used informally to write Kashmiri, particularly online. Only after the 8th Century A.D, the Kashmiri language was written in the Sharada script. However this script is not now widely used except for the religious ceremonies of the Kashmiri Pandits[12]. At present Kashmiri language is written in Perso-Arabic and Devanagari scripts (modified)[13]. Kashmiri is among one the languages written in Perso-Arabic script, which commonly indicates all vowel sounds. In general, Kashmiri Muslim community is more aligned with the Perso-Arabic script, whereas Devanagari script is being associated more with the Kashmiri Hindu community.

Kashmiri, though in the current scenario, has many dialects as well as scripts, but the speaker or user count is apparently declining. The authors are of the view that it is important to digitize resources for the Kashmiri language in Perso-Arabic script, the officially accepted as well as widely used script presented in Figure 2.

The Kashmiri language or Koshur is synonymous with the identity of Kashmiri people. It shows linguistic strata of different ages, such as Vedic, Buddhist Sanskrit, Pali, Kharoshthi Prakrit, etc.[14]. Dr. Grierson in 1919 said, "Kashmiri is the only Dardic languages that has a literature". Kashmiri literature is about 750 years old; this is the period of advent of literature in many

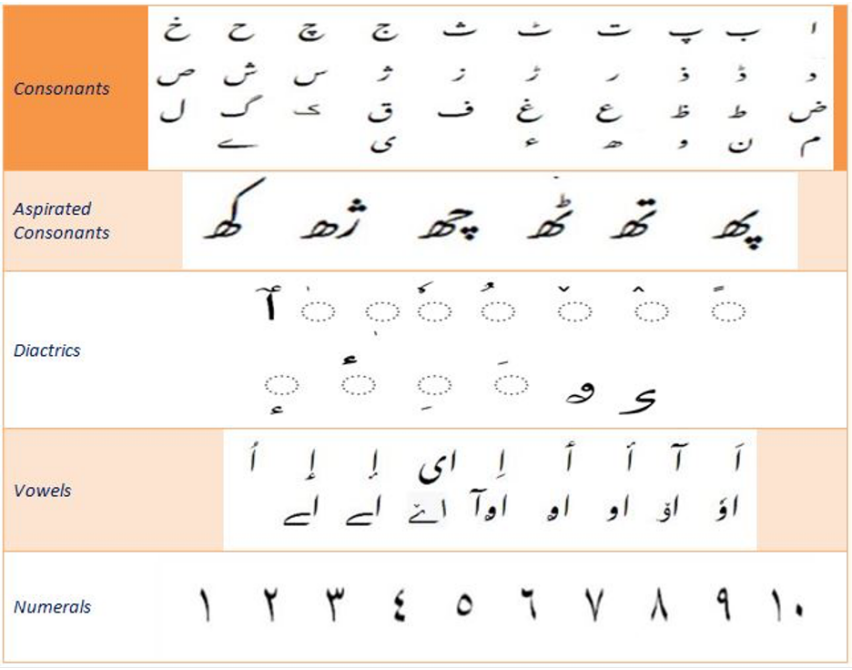

Fig 2. Perso-Arabic Script for the Kashmiri Language

modern languages such as English.

Kashmiri is a member of the Dardic subgroup (by its origin) of the Indo-Aryan language family (in character). The Kashmiri vocabulary is mixed, reflecting the past of the region, containing Dardic, Sanskrit, Punjabi, and Persian components. In vocabulary and alphabet collection, religious differences are also apparent. The important spoken dialects are Maraaz (southern and southeastern region), Kamraaz (northern and northwestern region), and Yamraaz (Srinagar and some of its surrounding areas) [15].

3 NLP Research Resources Available for the Kashmiri Language

Although the Kashmiri language is rich in terms of literature, poetry, religious manuscripts, etc. But there is scarcity of NLP resources for the Kashmiri languages, which are very essential to undertake any NLP related research in Kashmiri language. Currently, there is a dearth of digital resources available irrespective of the type of resource. The various resources & their profile is given as

3.1 Corpus

The ILCI - Indian Languages Corpora Initiative is a TDIL project to develop language corpora based on the national standard. Under this initiative, in the first phase development of parallel annotated corpora, India's national standard in POS annotation for 12 Languages including English took place, resulting in corpora creation of 600,000 sentences, with 16 words for each sentence. In phase second it included Nepali, Bodo, Assamese, Manipuri, and Kannada, and added 11, 00,000 new sentences for previous as well as newly added languages. The total size of the language corpora resulted in approximately 27 million parallel annotated and chunked words in the domain of Health and Tourism, Agriculture, and Entertainment [16]. But there is no mention of corpora for the Kashmiri language, whether it is parallel, Monolingual, or tagged except 256 English-Kashmiri translated strings. Under TDIL Indowordnet project for Indian languages created a Kashmiri word-net containing 29469 words, with each synset/word containing a POS category, an example of its use, and a set of synonyms [17,18].

A monolingual corpus consisting of about 2.3 million words [19] has been developed under the EMILLE Project (Enabling Minority Language Engineering) being supported by the Universities of Lancaster and Sheffield.

A raw textbook consisting of 466,054, extracted from different newspapers, magazines, and books has been recently prepared by the Central Institute of Indian Languages (CIIL) Mysore. It consists of the text mostly from the domains of Social sciences

and Aesthetics, etc.[20].

3.2 Other Resources

The other potential sources for creation of NLP resources include a news daily namely Sangarmal[21], published in Kashmiri Language (Perso-Arabic script). The paper however is being published in image format as an e-paper and there is no text file available which could be put in use for any data processing related operation. There is another Kashmiri Language daily Koshur Akhbar[22]. Further Soan Meeraas[23] is a Kashmiri language weekly that started its publication in 2006 from Srinagar with a mission to promote linguistic, cultural, and traditional virtues of the region besides sharing political views.

4 State of NLP Research for the Kashmiri Language

NLP specific research in Kashmiri language is monopolized by its resource-scarce nature. As there is a dearth of resources available pertaining to Kashmiri language, its research profile is thus very low. There is a wide research gap in this area and accordingly it is purely open to any research in terms of Natural Language Processing Phenomena.

NLP research in Kashmiri is mostly limited to the areas of (i) Machine Transliteration (ii) Parts of Speech (PoS) tagging (iii) Machine Translation. Given below is a brief account of the work done in these areas.

4.1 Machine Transliteration

There has been just one research report about Machine Transliteration involving Kashmiri as a participating Language[24]. The work presented in the paper uses a hybrid approach with direct mapping used for certain words that are in its knowledge base. Moreover, for direct mapping (word-to-word transliteration), phonemes are also generated from phonetic dictionaries, and rules are created to transliterate these phonemes. The grapheme of an English word is used to extract its phonemes. For certain words that are not subject to transliteration either by direct method or by rule generation a simple character to character, mapping is done.

4.2 Part-of-Speech (PoS) tagging

Part-of-speech (PoS) tagging (word-category-disambiguation/grammatical tagging) has emerged as an initial preprocessing step in the NLP process to determine the categories of morphemes used in a sentence. The elementary work done with regard to Kashmiri so far as pos-tagging is concerned is the development of a flat tagset in consonance with ILMT (Indian Language Machine Translation) guidelines under the project "Development of Language Tools and Linguistic Resources for Kashmiri" at the Department of the linguistics university of Kashmir. The ILMT tag set[25] has 21 parts of speech, and tag set for the Kashmiri language[26] is given in Table 1 .

Table 1. Tagset for Kashmiri Language

Category	Category
Noun	Quantifiers
Nloc	Cardinal
Noun with cases	Ordinal
Proper Noun	Intensifier
Pronoun	Interjection
Demonstratives	Negation
Verb Finite	Sym
Verb with cases	Compounds
Verb Aux	Reduplicative
Adjective	Echo
Adverb	Unknown
Postposition	Vocative
Particles	Equatives
Conjucts	Part of Word
Question Words	

A Model proposed POS-Tagging Kashmiri Language is based on probabilistic procedures using Conditional Random Filters (CRF). The pos-tagger was implemented with 30000 words. The POS-Tagger achieved an overall accuracy of about 80.64%. However, by increasing the training data the POS-Tagger may result in better performance as was evident by its result summary. The system performance got raised from 67.22% to 81.10% by varying the training data size from 15000 to 27000 [27,28].

4.3 Machine Translation

Translating English sentences to Kashmiri sentences has been presented in order to simple assertive sentences [29,30]. It consists of various phases such as analyzing morphology, part of speech(PoS) tagging, local word grouping, etc. to convert the simple assertive English sentences into Kashmiri sentences.

Kashmiri text to English text translation system reported in the literature uses a learning approach to machine translation. The work presented is based on Encoder-Decoder architecture. Long-short-term memory (LSTM) is used as the building block for both the encoder and decoder. The system has been evaluated against various automatic metrics and is having a good score across all of them. The efficiency metric is presented in Table 2. The paper is the first of its kind involving a parallel dataset [31].

Table 2. Automatic Metric Values

Sl. No	Metric	Coefficients
1	BLUE (Bi-lingual Evaluation Understudy)	0.517588
2	WER (Word Error Rate)	0.589522
3	NIST (National Institute of Standards and Technology)	4.760154
4	METEOR (Metric for Evaluation of Translation with Explicit Ordering)	0.393298

4.4 Audio Processing

There is one research report regarding the processing of the Kashmiri language in audio domain. The system is based on a Mel-spectrogram-based approach using convolutional neural networks for officially spoken languages of Jammu & Kashmir involving Kashmir and Ladakhi languages. The dataset used in the study contains six languages i.e. Kashmiri, Ladakhi, Urdu, English, Hindi, and Dogri. Initially, speech segments were converted into Mel-spectrograms by using inverse Fourier transformation to log of Fourier transformation of a time-domain signal, and at the backend, CNN serves as a classifier. Experiments were conducted on recorded speech, IIIT-H, and VoxForge. It found that an average accuracy of 100 % was achieved by running the model at 100 epochs [32].

4.5 Other Research

Parsing in NLP is the activity of determining the syntactic structure of a given sentence by taking into consideration its constituent words based on defining the Grammar of the Language. In the literature, there is mention of one report about Kashmiri Parser [33]. The model is using a hybrid chunkier for the Kashmiri language resulting in better performance in comparison to the CRF-based technique of chunking. The CRF-based chunker produced results with an overall accuracy of 91.41%, but using hybrid chunker accuracy touched 94.87%, showing a significant improvement of 3.44%.

Kashmiri being a resource-scarce language with very few NLP resources available, there have been articles regarding research work about the creation of a Dependency tree bank [34]. The reported treebank consists of 1,000 sentences with 17,462 words, annotated with part-of-speech (POS), chunk and dependency information. It has been manually annotated using the approach of Paninian Computational Grammar (PCG). It also reports on the basic pipeline consisting of tokenizing, stemming, POS tagging, and interdependent chunk parsing.

As there are problems or challenges associated with each research study, to overcome these challenges and to address such issues is always of great relevance and helps in formulation of the solutions. As far as the Kashmiri language is concerned some reports or papers are available with respect to the issues concerning the translation of the Kashmiri language to the English language [35].

As languages are inherently ambiguous, there is a need to devise a procedure that will confirm the meaning of the word based on the context as there are a lot of words that need to disambiguate sense. There is only one report regarding word sense disambiguation of the Kashmiri language [36].

As we are living in a society of multimedia information. There is a need to process such multimedia information to get the maximum out of it. The media may be in terms of text, image, audio, and, video, etc. Processing data from images is challenging

more so far cursive languages. The Kashmiri languages follow a cursive type of script so it is no exception to such challenges. Two papers involving the Kashmiri text processing in document images are published one dealing with the identification of the Kashmiri script using profile projections known as Peaks and valleys producing 96.2% accurate results and another involving exposing issues in document image processing for text identification [37,38].

5 Issues concerning Kashmiri as a language

The Kashmiri Language has lost its sheen in terms of its use in, formal as well as informal settings. There are multiple reasons for the same. One among them is the dominance of English (a universally used language) in non-verbal as well as verbal interactions, & dominance of Urdu which served as the official language for Jammu & Kashmir in the recent past as Kashmiri was not given the status of the Official Language as a result of which its witnessed a decline in its speaker cum user base. Another reasons is its limited or no use in most of the field, as a result of which it could not evolve in terms of addition of new words .

Although Kashmiri has flourished in ceremonial functions mostly marriages, folklore, theater, and other literary functions, the differences between the spoken and the written form of language are quite large.

6 Conclusion

The work presented in this paper is the first of its kind highlighting the current scenario of NLP resources cum research available for the Kashmiri Language. The conclusion of this paper consists of 3 parts as described below:

- As the Kashmiri is a highly low-resource language, nearly all the research work done in the field of NLP has been reported in this paper. There exists a single report of Machine Transliteration, Machine Translation, Parser, and Dependency treebank. However, there are also two reports about PoS-Tagging with good accuracy.
- There are three publically available datasets for non-commercial use with reference to the Kashmir language, one being Translated Strings. Wordnet which includes POS- Tag category, synonyms & their use through Example sentences. There also exists a monolingual corpus & three newspapers.
- There is strong commendation of having a good resource base for undertaking NLP specific research towards Kashmiri Language. Crowdsourcing as a means to create linguistic data need to be explored as has been done for many Indian Languages as a resource creation approach, particularly for generating data for machine Translation [39], as no such step has been reported with reference to Kashmiri as a participating Language.

References

1) Martinez REL, Sierra G. Research Trends in the International Literature on Natural Language Processing, 2000-2019 - A Bibliometric Study. *Journal of Scientometric Research*. 2020;9(3):310–318. Available from: https://doi.org/10.5530/jscires.9.3.38.
2) Gupta V, Joshi N, Mathur I. Advanced Machine Learning Techniques in Natural Language Processing for Indian Languages. *Smart Techniques for a Smarter Planet*. 2019;374:117–144. Available from: https://doi.org/10.1007/978-3-030-03131-2_7.
3) Languages of India. . Available from: https://en.wikipedia.org/wiki/Languages_of_India.
4) Rajan A, Salgaonkar A, Joshi R. A survey of Konkani NLP resources. *Computer Science Review*. 2020;38:100299. Available from: https://doi.org/10.1016/j.cosrev.2020.100299.
5) Kashmiri language. . Available from: https://en.wikipedia.org/wiki/Kashmiri_language.
6) Koshur: An Introduction to Spoken Kashmiri. . Available from: https://www.koshur.org/.
7) Snedden C. Understanding Kashmir and Kashmiris. *Choice Reviews Online*. 2016;53(08). Available from: http://dx.doi.org/10.5860/choice.195226.
8) Shakil M. The languages of Erstwhile State of. *Jammu & Kashmir Academia edu*. 2011. Available from: https://www.academia.edu/en/6485567/Languages_of_Erstwhile_State_of_Jammu_Kashmir_A_Preliminary_Study_.
9) Scheduled Languages of India. . Available from: https://en.wikipedia.org/wiki/Languages_of_India#Scheduled_languages.
10) The Jammu and Kashmir Official Languages Act, 2020, Ministry of Law and Justice (Legislative Department) New Delhi, The 27th September, 2020. 2020. Available from: https://prsindia.org/files/bills_acts/bills_parliament/2020/Jammu%20and%20Kashmir%20Official%20Languages%20Act,%202020.pdf.
11) Kashmiri made compulsory subject in schools. . Available from: https://www.oneindia.com/2008/11/01/kashmiri-made-compulsory-subject-in-schools-1225558278.html.
12) The Sharada Script: Origin and Development. . Available from: http://www.koausa.org/Languages/Sharda.html.
13) Kashmiri Alphabet, Pronouncation & Language. . Available from: https://omniglot.com/writing/kashmiri.htm.
14) Kashmiri Language: Roots, Evolution and Affinity. . Available from: http://www.koausa.org/Languages/Shashi.html.
15) An Introduction To Spoken Kashmiri. . Available from: http://www.koshur.org/SpokenKashmiri/Introduction/index.html.
16) Indian Languages Corpora Initiative. . Available from: http://sanskrit.jnu.ac.in/projects/ilci.jsp?proj=ilci.
17) TDIL: Technology Development for Indian Languages Programme. India. . Available from: https://tdil.meity.gov.in/.
18) Dash NS, Bhattacharyya P, Pawar JD. The WordNet in Indian Languages. Springer Singapore. 2017. Available from: https://doi.org/10.1007/978-981-10-1909-8.
19) Baker P, Hardie A, Mcenery T, Cunningham H, Gaizauskas R. EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. 2002;p. 219–225.

20) Ramamoorthy L, Choudhary N, Bhat SM, Central Institute of Indian Languages, Mysore. A Gold Standard Kashmiri Raw Text Corpus. Mysore. 2019. Available from: https://data.ldcil.org/a-gold-standard-kashmiri-raw-text-corpus.
21) Sangarmal Epaper. . Available from: http://epaperdaily.com/indian/sangarmal-kashmir-jak.html.
22) Koshur Akhbar. 2022. Available from: http://www.epaper-hub.com/india/epaper/koshur-akhbar.aspx.
23) Soanmeeraas News Weekly. . Available from: https://www.soanmeeraas.in/e-paper.
24) Aadil M, Asger M. English to Kashmiri Transliteration System - A Hybrid Approach. *International Journal of Computer Applications*. 2017;162(12):5–8. Available from: https://doi.org/10.5120/ijca2017913418.
25) Bharati A, Sharma DM, Bai L, Sangal R, Corra A. Annotating Corpora Guidelines For POS And Chunk. IIIT, Hyderabad: Language Technologies Research Centre. . Available from: https://www.sketchengine.eu/wp-content/uploads/posguidelines_indian_languages.pdf.
26) Kak AA, Mehdi N, Lawaye A. What should be and What should not be? Developing a POS tagset for Kashmiri. *Interdisciplinary Journal of Linguistics (IJL)*. 2009;2:185–96.
27) Lawaye AA, Purkayastha BS. Kashmir Part of Speech Tagger Using CRF. *Paripex - Indian Journal Of Research*. 2012;3(3):37–38. Available from: http://doi.org/10.15373/22501991/mar2014/11.
28) Lawaye AA, Purkayastha S. Towards Developing a Hierarchical Part of Speech Tagger for Kashmiri: Hybrid Approach. *Proceedings of the 2nd National Conference on Advancement in the Era of Multidisciplinary Systems (AEMDS-2013)*. 2013;p. 187–192. Available from: https://www.researchgate.net/publication/321937581_Towards_Developing_a_Hierarchical_Part_of_Speech_Tagger_for_Kashmiri_Hybrid_Approach.
29) Mustaif I, Bansal K. English to Kashmiri Translation System: Using Example Based Machine Translation. *International Journal of Innovative Research in Computer and Communication Engineering*. 2015;2015(10):193–194. Available from: https://1library.net/document/y80x1w5q-english-kashmiri-translation-using-example-machine-translation-approach.html.
30) Mustaif I, Bansal K. English to Kashmiri Translation System. *International Journal of Advanced Research in Computer Science and Technology*. 2015;2015(2):193–194. Available from: http://ijarcst.com/doc/vol3issue2/ver2/mutasif.pdf.
31) Lone NA, Giri KJ, Bashir R. Machine Intelligence for Language Translation from Kashmiri to English. *Journal of Information & Knowledge Management*. 2022. Available from: https://doi.org/10.1142/S0219649222500745.
32) Thukroo IA, Bashir R. Spoken Language Identification System for Kashmiri and Related Languages Using Mel-Spectrograms and Deep Learning Approach. *2021 7th International Conference on Signal Processing and Communication (ICSC)*. 2021;p. 250–255. Available from: https://ieeexplore.ieee.org/document/9673212/.
33) Bhat RA, Sharma DM. A hybrid approach to kashmiri shallow parsing. InThe 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. 2011. Available from: https://www.yumpu.com/en/document/view/12199664/a-hybrid-approach-to-kashmiri-shallow-parsing-iiit-hyderabad.
34) Bhat RA, Bhat SM, Sharma DM. Towards building a Kashmiri treebank: setting up the annotation pipeline. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2014;p. 748–752.
35) Lone NA, Giri KJ, Bashir R. Issues in Machine Translation—A Case Study of the Kashmiri Language. In: Machine Intelligence and Data Science Applications. Springer Nature Singapore. 2022;p. 117–123. Available from: https://link.springer.com/chapter/10.1007/978-981-19-2347-0_10.
36) Mir AT, Lawaye AA. Word Sense Disambiguation for Kashmiri Language Using Supervised Machine Learning. *Proceedings of the 17th International Conference on Natural Language Processing*. 2021;p. 243–245.
37) Lone NA, Giri KJ, Bashir R, R. Processing Kashmiri Text from a Document Image and Associated Challenges. In: In Proceedings of the 6th International Conference on Advance Computing and Intelligent Engineering 2023. Springer. 2023;p. 449–455.
38) Bashir R, Quadri S. Identification of Kashmiri script in a bilingual document image. *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*. 2013;p. 575–579. Available from: https://doi.org/10.1109/ICIIP.2013.6707658.
39) Kunchukuttan A, Roy S, Patel P, Ladha K, Gupta S, Khapra MM, et al. Experiences in resource generation for machine translation through crowdsourcing. *InProceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. 2012;p. 384–391. Available from: https://aclanthology.org/L12-1127/.