# Dr. Atul Kumar Ojha

Panlingua Language Processing LLP
#20 U Ground Floor Road No 23,
East Punjabi Bagh,
New Delhi-110028, India
+91-7838979219
shashwatup9k@gmail.com
Skype@shashwat9k
Google Scholar: https://tinyurl.com/y4m6pu82

## RESEARCH INTERESTS

Statistical & Neural Machine Translation • Automatic metrics for MT evaluation • Corpus Mining • Machine Learning • Hate and Aggressive Speech • Artificial Intelligence •Big data & Computational Linguistics

## PROFESSIONAL EXPERIENCE

**Co-founded Panlingua Language Processing LLP**
(July'17 - present)

**Sr. NLP Research Engineer** 01/08/2017 – 31/07/2018
*Machine Translation Evaluation Platform*
J.N.U, New Delhi

Development of GUI-based MT tool
R&D on evaluation metrics

**Sr. Linguist** 11/11/2014-20/01/2017
*Indian Languages Corpora Initiative (ILCI)*
J.N.U, New Delhi

Project Management
Linguistic Training
Testing project tools
Assist in creation of GUI-based tool

**Linguist** 01/08/2013 - 10/11/2014
*Indian Languages Corpora Initiative (ILCI*
J.N.U, New Delhi

POS annotation
Validation of annotated data

**Linguist** 31/12/2012 - 01/03/2013
*Short Term Goal Oriented Project-Hindi*,
LDC-IL, CIIL, Mysore

Speech annotation

**Language Editor** 25/04/ 2011 - 21/10/2011
*Development of Sanskrit Computational Toolkit and Sanskrit-Hindi Machine Translation System (SHMT)*

Dept. of Sanskrit studies, University of Hyderabad

Evaluation of SHMT system and linguistic resolutions

## PROJECTS

- Co-ordinator in the project titled "Implementation of Indian languages in the IMAGACT Multilingual Ontology of Action" under the University of Florence-JNU joint collaboration.
- Collaborator in the project titled "Automatic detection of verbal threat in Hindi and English aggressive speech" under UGC-UKIERI Thematic Partnerships 2014**.**

## TEACHING EXPERIENCE/INVITED TALKS

- Teaching assistant for 'Machine Learning and Statistical Machine Translation' to M.A. students with Prof. Girish Nath Jha, JNU, New Delhi from 2014-present.
- Lectures in "Workshop on Computational Linguistics & Machine Translation" from 19-29 September 2017 at JNU, New Delhi.
- Lectures in "Training of Computational Linguistics" from 02-12 August 2016 at Yogyakarta State University, Yogyakarta, Indonesia.

## EDUCATIONAL QUALIFICATION

- PhD (awarded date is: March 29, 2019) in the Natural Language Processing / Computational Linguistics, titled "**English-Bhojpuri SMT System: Insights from the Kāraka Model"** from SSIS, JNU, New Delhi.
- M.Phil in Hindi-Language Technology at M.G.A. Hindi University in 2010-11 with I division (7.09 FGPA).
- M.A. in Linguistics at Banaras Hindu University in 2010 with I division (6.94 DGPA/CGPA).
- B.A. in Sanskrit, Ancient History & English Literature at University of Allahabad in 2008 with II division.

## OTHER QUALIFICATIONS

- Completed 'XML Technologies and Semantic Web' and 'Machine translation and parallel corpora' courses during the exchange student at University of Zürich, Switzerland.
- Completed 'Machine Learning' course during the PhD coursework from School of Computer Science, JNU, New Delhi

## TECHNICAL SKILLS

- Programming/Scripting Languages: Java, Python and Shell Scripting
- Web Technologies: HTML, JSP and XML
- NLP tools: Moses, OpenNMT, NLTK, Apertuim, SVM, CRF, Praat

## DEVELOPMENT EXPERIENCE

- GUI of Moses based Statistical Machine Translator system under the MTEP project
- For R& D trained statistical and neural-based MT systems for Indian languages.

- SVM-based Hindi PoS Tagger for ILCI project
- CRF-based Bhojpuri PoS Tagger for research work
- Hindi and Indian-English Chunker for ILCI project
- Machine-readable multilingual dictionary for Bhojpuri-Hindi-English

## Published Papers

- ***Panlingua-KMI SMT and NMT System @ WMT Similar Language Translation 2019*** (with Ritesh Kumar, Akanksha Bansal and Priya Rani) submitted in WMT Similar Language Translation Task under the WMT Shared Task 2019.
- ***KMI-Coling at Semeval-2019 Task 6: Exploring N-grams for Offensive Language Detection*** (with Priya Rani) under publication in OffensEval 2019 under the International Workshop on Semantic Evaluation 2019, NAACL-2019.
- ***Issues & Challenges in Building SMT Systems for Lesser-known Languages: The Case of English-Bhojpuri & English-Garhwali*** (with Arushi Uniyal and Grish Nath Jha) accepted for publication in Proceedings of the First International Sanskrit and Other Indian Languages Technology (SOIL-Tech), February 15-17, 2019, New Delhi, India.
- ***Aggression in Hindi & English Speech: Acoustic Correlates & Automatic Identification*** (with Ritesh Kumar, Bornini Lahiri and Chingrimung Lungleng) accepted for publication in Proceedings of the First International Sanskrit and Other Indian Languages Technology (SOIL-Tech), February 15-17, 2019, New Delhi, India.
- ***The RGNLP Machine Translation Systems for WAT 2018*** (with Koel Dutta Chowdhury, Chao-Hong Liu, Karan Saxena) in Proceedings of the 5th Workshop on Asian Translation (WAT2018) under the PACLIC-32 Conference 2018, Hong Kong, China.
- ***Benchmarking Aggression Identification in Social Media*** (with Ritesh Kumar, Shervin Malmasi, and Marcos Zampieri) in Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC-1) under the COLIN2018, Santa Fe, USA.
- ***A corpus based study of semantics of bare nominals in Magahi and Bhojpuri: The case of article-less languages*** (with Deepak Alok and Sriniket Kumar Mishra) in Linguistic Ecology of Bihar, Europa Publisher (in press).
- ***Automatic Identification of Closely-related Indian Languages: Resources and Experiments*** (with Ritesh Kumar, Bornini Lahiri, Deepak Alok, Mayank Jain, Abdul Basit, Yogesh Dawar)
- ***Automatic Language Identification System for Hindi and Magahi*** (with Priya Rani and Girish Nath Jha) in proceedings of the 4th Workshop on Indian Language Data: Resources and Evaluation (under the 11th LREC2018, May 07-12, 2018) by European Language Resources Association (ISBN: 979-10-95546-09-2 EAN: 9791095546092).
- ***Challenges in Annotation and Domain Adaptation in Hindi POS Tagger: with Reference to Cricket*** (with Anupama Pandey, Sirshti Singh, and Girish Nath Jha) in the proceedings of ICATCCT - 2017 (IEEE ISBN: 978-1-5090-6036-8).
- ***The IMAGACT4ALL Ontology of Animated Images: Implications for Theoretical and Machine Translation of Action Verbs from English-Indian Languages*** ( with Pitambar Behera, Sharmin Muzaffar, and Girish Nath Jha) in the proceeding of 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP-2016) under the COLING2016.
- ***A Hybrid Chunker for Hindi and Indian English*** (with Pitambar Behera, Srishti Singh, and Girish Nath Jha) in the Proceedings of 3rd Workshop on Indian the Language Data: Resources and Evaluation, 10th International Conference on Language Resources and

Evaluation (LREC-2016), Portorož (Slovenia), 2016.

- ***Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri*** (co-authored with Pitambar Behera, Srishti Singh, and Girish Nath Jha) in the proceeding of 7th LTC 2015 Conference during November 27-29, 2015, Poznan, Poland.
- ***Issues and Challenges in Developing Statistical PoS Taggers for Sambalpuri*** (co-authored with Pitambar Behera, and Girish Nath Jha) in the proceeding of 7th LTC 2015 Conference during November 27-29, 2015, Poznan, Poland.
- ***Evaluation on Hindi-English MT Systems*** (co-authored with Akansha Bansal, Sumedh Hadke & Girish Nath Jha) in the Proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation, 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 2014.
- ***Mapping Indian Languages onto the IMAGACT Visual Ontology of Action*** (co-authored with Massimo Moneglia, Susan W. Brown, Aniruddha Kar, Anand Kumar,Heliana Mello, Niharika, Girish Nath Jha, Bhaskar Ray, & Annu Sharma) in the Proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation, 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 2014.
- ***A Language Engineering Approach to Ameliorate Hindi Morph Analyzer*** (co-authored with Sanket Pathak, & Rashid Ahmad) in the Shodh Prerak, A Multidisciplinary Quarterly International Refereed Research Journal, ISSN2231-413X, Vol. No.-IV, Issue -4 December 2012.
- ***A Language Engineering Approach to Enhance the Accuracy of Machine Translation Systems*** (co-authored with Sanket Pathak & Rashid Ahmad) in the Shodh Prerak, A Multidisciplinary Quarterly International Refereed Research Journal, ISSN2231-413X, Vol. No.-I, Issue -1, January 2012.

## PAPERS PRESENTED

- Presented a paper on **"Developing English-Javanese Statistical Machine Translation System"** (co-authored with Girish Nath Jha) in International Conference on India & Southeast Asia: One Indic Belt, Shared Culture & Common Destiny, JNU, New Delhi, 26-28 April 2018.
- Presented a paper on **"Indo Aryan languages on IMAGACT"**(co-authored with Girish Nath Jha, Sharmin Muzaffar and Pitambar Behera) in IMAGACT Panel, MODELACT Conference on Action, Language and Cognition, CNR, Rome, 6-7 June 2016.
- Presented a paper on ***"Developing a Machine Readable Multilingual Dictionary for Bhojpuri-Hindi-English***" in ELKL-4 2016 Conference during February 25-27, 2016, Agra, India.
- Presented a paper on "***Problem areas in the Performance of the Hindi Shallow Parser***" (co-authored with Arushi Uniyal) in 8th SCONLI, at Department of Linguistics, University of Kashmir during March 24-26, 2014.
- Presented a paper on "***Mirjapuri: A branch of Bhojpuri Tree***," (co-authored with Shailendra Kumar) in Endangered and Lesser Known Languages: Challenges and Responses, at Department of Linguistics, Luck now University during October 11-13, 2012.
- Presented a paper on "***Syntax and Semantics of the Postpositions 'ke' in Bhojpuri,"*** (co-authored with Gayetri Thakur) in 33AICL, at Chandigarh during, October 1-3, 2011.

- Presented a paper on “***Hindi aur Sanskrit nipatom ka tulanãtmak adhyayan*”** ***(A Comparative study of Hindi and Sanskrit Particles)*** at XXXIV Indian Social Science Congress on 27th to 31st Dec 2010 organized by Guwahati University.

## WORKSHOPS/SEMINARS/SYMPOSIUMS/CONFERENCES ATTENDED

- Participated in "8th Lisbon Machine Learning School" during 14th-21st June 2018 at Instituto Superior Técnico (IST), Lisbon, Portugal.
- Participated in “Workshop on Neural Machine Translation” during 04-09th December 2017 at IIT-Patna.
- Participated in "Machine Translation meets Translators" workshop during 15th-17th May 2017 at University of Zürich, Switzerland.
- Participated in "Summer School of RBMT 2016 at University of Alicante, Spain. Participated in "Winter School on Speech and Audio Processing (WiSSAP) 2016 at SSN, Chennai.
- Participated in "Winter School on Speech and Audio Processing (WiSSAP) 2015 at DAICT, Gandhinagar.
- Participated in "Winter School on Speech and Audio Processing (WiSSAP) 2014 at IIIT-Hyderabad.
- Participated in “Short Training program in the Moses statistical MT platform" with Hieu Hoang as the resource person during December 25-29, 2014 in the Computational Linguistics Lab at the Special Centre for Sanskrit Studies, JNU, India.
- Participated in “workshop on Advances in Multimodal Multilingual Translation Process" on Nov. 28, 2013 at JSSATE Noida, India.
- Participated in “Workshop on Multilingual Technologies" at Department of Computer Science, Punjabi University, Patiala, India on 11-17 November, 2013.

## ORGANIZED/MANAGED WORKSHOPS

- Organising Shared Task on Linguistically Motivated for LoRes languages under the 2 LoRes Workshop as co-organiser under the MT Summit XVII-2019 at Dublin, Ireland.
- Organising 2 LoRes Workshop as co-organiser under the MT Summit XVII-2019 at Dublin, Ireland.
- Organising Low-level NLP Tools for Magahi and Bhojpuri Shared Task under the NSURL-2019.
- Managed First International Conference on Sanskit & Other Indian Language Technology-2019 at JNU, New Delhi
- Organized First Shared Task on Aggression Identification under the on First Workshop on Trolling, Aggression and Cyberbullying (TRAC-1) @ COLING 2018
- Organized First Workshop on Trolling, Aggression and Cyberbullying (TRAC-1) @ COLING 2018
- Managed 4th Workshop on Indian Language Data: Resources and Evaluation (WILDRE-4) under the LREC2018, Miyazaki (Japan), May 12, 2018
- Managed 3th Workshop on Indian Language Data: Resources and Evaluation (WILDRE-3) under the LREC2016, Portorož (Slovenia), May 24, 2016
- Managed Science & Technology for Sanskrit at the Special Center for Sanskrit Studies, March 10-12, 2014
- Managed Sharada workshop-3 at the Special Center for Sanskrit Studies, Feb 18-20, 2014
- Managed Sharada workshop-2 at the Special Center for Sanskrit Studies, Nov 8-10, 2013

## AWARDS

- Received Google travel fellowship to participate in 8th Lisbon Machine Learning School (LxMLS 2018) from 14-21st June 2018 at Instituto Superior Técnico (IST), Lisbon, Portugal
- Received Heyning-Roelli Foundation fellowship to participate as an exchange student from 1st Feb- 30th June 2017 at University of Zürich, Switzerland
- Received travel grant from JNU, New Delhi to present paper in LREC-2016, Slovenia
- Received travel grant from Microsoft Research India to present paper in LREC-2014, Iceland

## EDITED PROCEEDINGS/BOOKS

- Book on ***Veda As Global Heritage Scientific Perspective*** (with Girish Nath Jha, Sudhir Arya, Abhijit Dixit), Vidya Nidhi Publisher (ISBN-938553969-8).
- ***First Workshop on Trolling, Aggression and Cyberbullying*** (TRAC-2018) proceedings under the 27th COLING2018 (with Ritesh Kumar, Marcos Zampieri, Shervin Malmasi).
- 4th Workshop on ***Indian Language Data Resource and Evaluation*** (WILDRE-4) proceedings under the LREC 2018 (with Girish Nath Jha, Kalika Bali and Sobha L.)
- 3rd Workshop on ***Indian Language Data Resource and Evaluation*** (WILDRE-3) proceedings under the LREC2016 (with Girish Nath Jha, Kalika Bali and Sobha L.)

## COMMITTEE MEMBER AND REVIEWER FOR CONFERENCES/WORKSHOPS

- **Reviewer:** TRAC-1, 2018, CoNLL 2018 UD Shared Task, ICON-2018
- **Program committee/Organiser member:** WILDRE-3, WILDRE-4, TRAC-1, 2018, CoNLL 2018 UD Shared Task**,** SOIL-Tech 2019, 2LoRes 2019 Workshop under the MT Summit XVII-2019

## LANGUAGES KNOWN

Bhojpuri | Hindi | English | Sanskrit

Akanksha Bansal

akanksha.bansal15@gmail.com
+91-9560774110
Delhi, India

## SUMMARY

I’ve been practicing linguistics as a researcher, academician and a professional for approximately 10 years. While my education and teaching experience strengthened the theoretical base in all structural aspects of language and linguistics, my professional practice helped me develop a keen eye for quality and enhanced my analytical skills. My professional roles have ranged from that of a phonetician to a computational linguist, allowing me to gain a hands-on experience in transcription, transliteration, annotation, review and validation, research, analysis, documentation and training. I have worked on a vast array of data sets and operated several data management tools. My experience in project management not only developed my nuances as a team player but also the skill to identify tools and mechanisms to enhance the speed and quality of any process.

## WORK EXPERIENCE

**Co-founded Panlingua Language Processing LLP**
**(July’17 - present)**

**Freelance Consultant for Language and Linguistics**
**(July’17- present)**

Linguistic data management and analysis for Hindi and Indian English
Developing grammar-based rules and optimising phone-sets

**Learning Scientist** — ***Liqvid English Edge Pvt. Ltd.***
**Feb’16 - Jun’17** — ***NOIDA***

- *Assessment Design, Curriculum Design, and Learning Design*
- *Product Development and Management*
- *Training - Face to Face and Satellite-based*
- *Developing NLP features for eLearning*
- *Dealing with vendors for translation and quality analysis for product localisation in 9 indian languages.*

**Sr. Linguist** (Feb’13 - Aug’14) — ***Indian Languages Corpora Initiative (ILCI)***
**Jr. Linguist** (May’10 - Jul’12) — ***DeiTY sponsored Project (GOI)***

- *Corpus collection, domain classification, and source identification.*
- *Translation, Part-Of-Speech Annotation, Chunking (HIndi and English)*
- *Developing translation and annotation guidelines for Indian Languages*

- *Staff Training for corpora collection in other Indian Languages*
- *Maintaining common translation and annotation standards across 17 languages in India.*

## TEACHING EXPERIENCE

*(Responsible for Instruction, Curriculum Design, preparation of Readings List and Evaluation Criteria)*

| | |
|---|---|
| 2016 (Jan –Apr) | **Assistant Professor** (Guest) for ***English*** at Motilal Nehru College (Evening Studies), Delhi University. |
| 2015 (Jul – Dec) | **Assistant Professor** (Guest) for ***English Language*** at Shaheed Rajguru College of Applied Sciences for Women, Delhi University. |
| 2015 (Jul – Sept) | **Visiting Faculty** (Post-Graduation) at Pearl Design Academy for ***Communication Skills in English*** |
| 2012 – 2013 | **Teaching Assistant** at Jawaharlal Nehru University for courses in ***Linguistics*** |
| 2008 (May - Oct) | **IELTS Instructor** at Touchstone Educationals, Chandigarh |

## RESEARCH EXPERIENCE

- Ph.D. (2011 onwards) on "**Nonverbal Aspects of Communication in Social Media with reference to Technology based Transformations in Communication and Community**"

- MPhil (2009-10) on "**A Pragmatic study of Cyber Communication with special reference to the Multilingual Indian Context**"

## RESEARCH PAPERS

- Atul Kr., Akanksha Bansal, Sumedh Hadke & Girish Nath Jha, ***Evaluation on Hindi-English MT Systems*** in the Proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation, 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 2014. (ISBN 978-2-9517408-8-4 EAN 9782951740884)

- ***Linking and Referencing Multi-lingual corpora in Indian languages***, Esha Banerjee, Shiv Kaushik, Pinkey Nainwani, Akanksha Bansal, Girish Nath Jha in Human Language Technologies as a Challenge for Computer Science and Linguistics (proceedings of the 6th LTC), Zygmunt Vetulani & Hans Uszkoreit (ed), pp 65-68, Fundacja, Uniwersytetu im. A. Mickiewicza, Poznan, Poland, 2013

- ***Corpora creation for Indian Language Technologies – the ILCI project***, Akanksha Bansal, Esha Banerjee, Girish Nath Jha in Human Language Technologies as a Challenge for Computer Science and Linguistics (proceedings of the 6th LTC), Zygmunt Vetulani & Hans Uszkoreit (ed), pp 253-257, Fundacja, Uniwersytetu im. A. Mickiewicza, Poznan, Poland, 2013

- ***Issues in chunking parallel corpora: mapping Hindi-English verb group in ILCI***, Esha Banerjee, Akanksha Bansal, Girish Nath Jha, In Proceedings of 2nd Workshop on Indian Language Data:

Resources and Evaluation (WILDRE-2), Ninth International Conference on Language Resources and Evaluation, LREC 2014. (ISBN 978-2-9517408-8-4)

- Paper titled ***Encoding of Nonverbal Cues in Facebook*** presented at SALA, University of Hyderabad, Hyderabad, Feb 2014

- Presented a Poster titled ***An Overview of Prosodic Aspects in Cyber Communication: An Indian Perspective*** in The International Seminar on Prosodic Interface (ISPI-11) held at Jawaharlal Nehru University, New Delhi, 25-26 November 2011

## DEMOS AND WORKSHOPS

- Panlingua-KMI SMT and NMT System @ WMT Similar Language Translation 2019 (with Atul Ojha, Ritesh Kumar and Priya Rani) submitted in WMT Similar Language Translation Task under the WMT Shared Task 2019.

- Aggression, Abuse and Hyperpartisanship: Recognising a Multi-headed Hydra in the Cyberspace, a demo presented at ICTFLING 2019 held at IIT Patna from 11-12 January 2019.

## TECHNICAL KNOWLEDGE

***Languages****:* HTML, CSS, XML, Python
***Tools****:* Praat
***Applications***: MS-Office, Corel Draw, Adobe Photoshop

***OS***: Microsoft Windows, Linux, OS X
***Notations***: IPA, XSAMPA

## EDUCATIONAL QUALIFICATIONS

Cleared UGC-NET (December 2013) in English

| **Qualification** | **Subjects** | **Institute** |
|---|---|---|
| PhD (pursuing) | Linguistics | JNU, New Delhi |
| M.Phil (2009-11) | Linguistics | JNU, New Delhi |
| M.A. (2007-09) | English Literature | Panjab University, Chandigarh |
| B.A. Eng Hons (2004-07) | Psychology, Functional English, Elective English, English Honors | GCG-11; Panjab University, Chandigarh |

## LANGUAGES KNOWN

Hindi (Native), English, Punjabi