PaCMan: Parallel Corpora Management Workbench

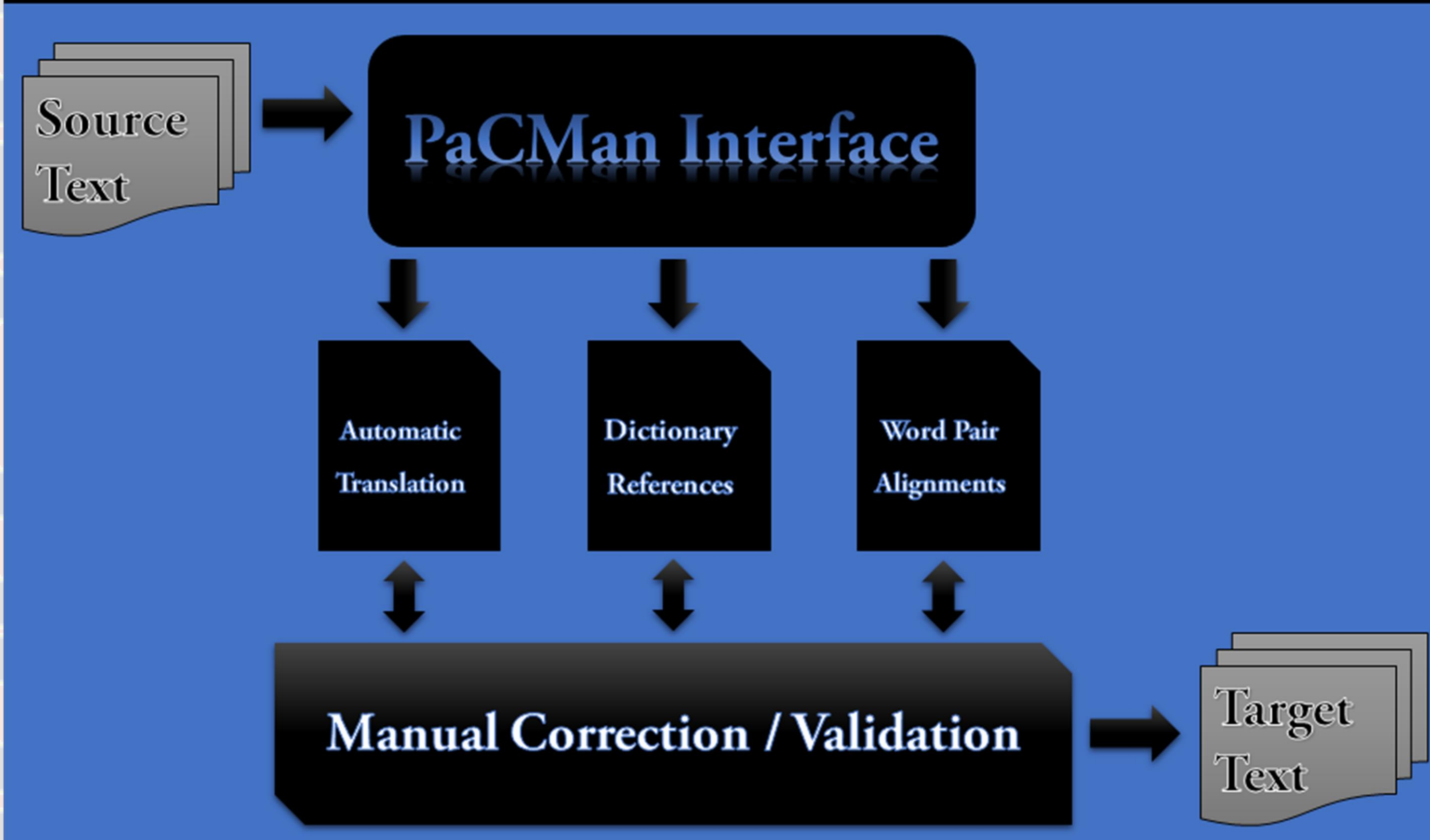Diptesh Kanojia   Manish Shrivastava   Raj Dabre   Pushpak Bhattacharyya
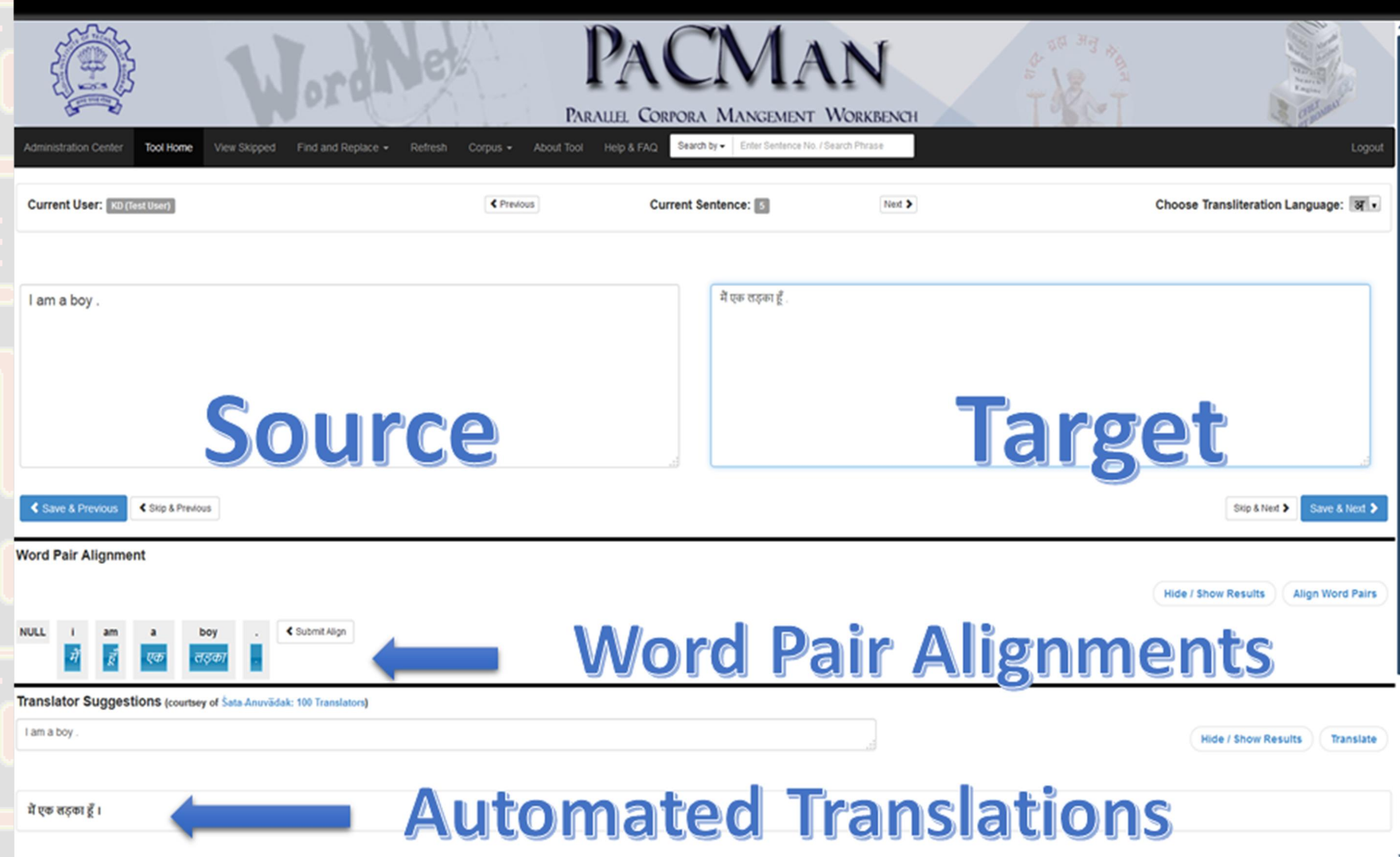
HTTP://WWW.CFILT.IITB.AC.IN/CORPORAMT/

Salient Features of PaCMan

- PaCMan aids Machine Translation
- Parallel Corpora Generation / Validation
- User / Corpora Management
- Task Assignment / Tracking
- Helps perfect Phrase Table generation

Workflow of the System
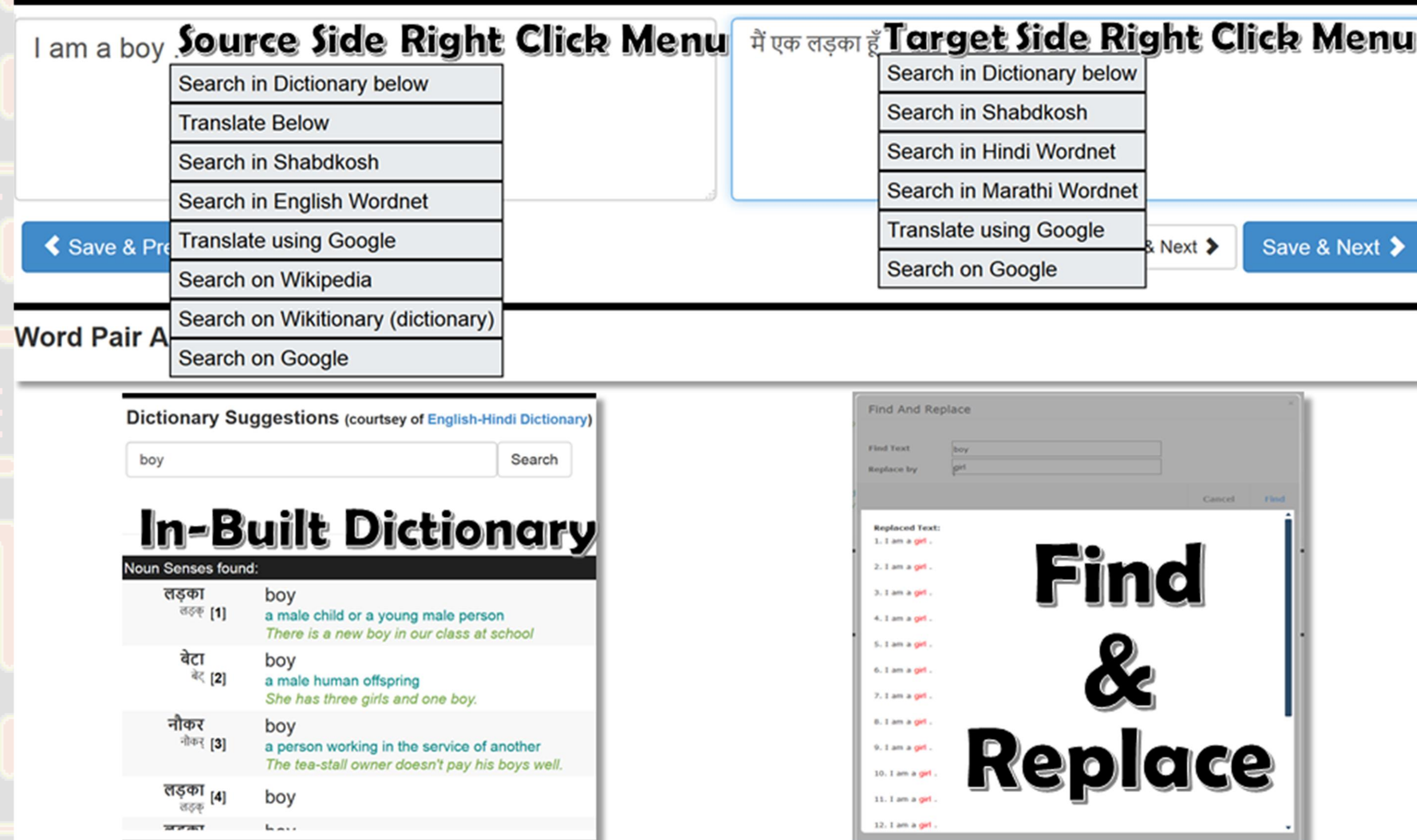

The one with The Screenshot

Lexicographer Aid

- In built Machine Translation system[1]
- References – a right click away!
  - Wikipedia / Wiktionary
  - WordNets / Google Translate / Shabdkosh
- Multilingual Phonetic Transliteration from Google APIs
- In built Hindi – English Dictionary

[1]Sata-Anuvadak: 100 Indic Translators

Various Glimpses

Tool Attributes

- Input - Simple text, Zip Archive of text files
- Output – Text Files, Database storage
- Saves lexicographer's session
- Search through Sentence ID / Word Phrase
- Find and Replace facility for the current session
- Allows skipping of sentences to be translated or validated later

The tool uses HTML/CSS, PHP, MySQL, Javascript, jQuery, Python

References

- Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-Anuvadak: Tackling multiway translation of Indian languages. In *Proceedings of the 9th Edition of its Language Resources and Evaluation Conference.*, LREC '14.
- Pushpak Bhattacharyya. 2006. English - Hindi dictionary. http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/.
- Prabhakar Pande, Dipak Narayan, Debasri Chakrabarti and Pushpak Bhattacharyya. 2002. An experience in building the IndoWordnet - a Wordnet for Hindi. In *Proceedings of First International Conference on Global WordNet*, Mysore, India.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.
- Najeh Hajlaoui and Christian Boitet. 2004. Polyphraz: A tool for the management of parallel corpora. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, MLR '04, pages 109–116, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yasu Imao. 2008. Casualconc. http://sites.google.com/site/casualconc/.
- Kilgray. 2006. memoq [computer software]. http://kilgray.com/products/memoq.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondˇrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51. SDL. 2007.
- SDL Trados studio [computer software]. http://www.translationzone.com/products/sdl-trados-studio/.
- Manish Soni. 2003. Shabdkosh. http://www.shabdkosh.com.
- Serge Yablonsky. 2003. The corpora management system based on java and oracle technologies. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 179–182, Stroudsburg, PA, USA. Association for Computational Linguistics