References

[1] Allen J. (1995). Natural Language Understanding. Benjamin/Cummings Publishing Company.

[2] Antony P.J. (2013). Machine Translation Approaches and Survey for Indian Languages. Computational Linguistics and Chinese Language Processing. Vol. 18, No. 1, March 2013, pp. 47-78.

[3] Antony P.J. and Soman, K.P. (2011). Parts Of Speech Tagging for Indian Languages: A Literature Survey. International Journal of Computer Applications (0975 – 8887).Volume 34– No.8, November 2011.

[4] Arulmozhi P., Rao R. K. and Sobha L. (2006). A Hybrid POS Tagger for a Relatively Free Word Order Language. In Proceedings of the Modeling and Shallow Parsing of Indian Language (MSPIL), Bombay. 79-85.

[5] Baskaran S. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information.

[6] Baskaran S. et al. (2007). Framework for a Common Parts-of-Speech Tagset for Indic Languages. http://research.microsoft.com/~baskaran/POSTagset/.

[7] Baskaran S. et al. (2008). Designing a Common POS-Tagset Framework for Indian Languages. The 6th Workshop on Asian Language Resources, 2008.

[8] Bharati A., Chaitanya V. and Sangal R. (1995). Natural Language Processing: A Paninian Perspective. Prentice Hall India.

[9] Brants T. (2000). TnT – A statistical part-of-sppech tagger. In Proceedings of the 6th Applied NLP Conference. 224-231.

[10] Brill E. (1992). A simple rule-based part of speech tagger. In Proceeding of third Conference on Applied Natural Language Processing, ACL, Trento, Italy.

[11] Brill E. (1994). Some Advances in Transformation-Based Part of speech tagging. In the Proceeding of Twelfth National Conference on Artificial Intelligence (AAAI-94).

[12] Brill E. (1995). Unsupervised Learning of Disambiguation Rules for Part of speech tagging. Natural Language Processing Using Very Large Corpora Text, Speech and Language Technology Volume 11, 1999, pp 27-42.

[13] Burange S., Devlakar, S. and Bhattacharyya, P. (2006). Rule Governed Marathi POS Tagging. In Proceeding of MSPIL, IIT Bombay, pp 69- 78.

[14] Chakrabarty A. et al. (2010). Experiences in building the Nepali WordNet - insights and challenges. In Proceedings of the 5th International Conference of the Global WordNet Association, IIT-Bombay, India.

[15] Chanod J. and Tapanainen P. (1995). Tagging French: comparing a statistical and a constraint-based method. In Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics, Dublin, Ireland.

[16] Chelliah S.L. (1997). A Grammar of Meithei. Mouton de Gruyter, Berlin.

[17] Church K.W. (1988). A stochastic parts program and noun phrase for unrestricted text. Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics, Dublin, Ireland.

[18] Choudhury S. I., Singh, L. S., Borgohain, S. and Das, P.K. (2004). Morphological Analyzer for Manipuri: Design and Implementation. In Proceedings of AACC, Kathmandu, Nepal, pp 123-129.

[19] Crystal D. (1980). A First Dictionary of Linguistics and Phonetics. Colorado: Westview Press Boulder.

[20] Cutting D., Kupiec J, Pederson J. and Sibun P. (1992). A practical part-of-speech tagger. In Proceedings of the 3rd Conference on Applied NLP. 133-140.

[21] Daelemans W. and Antal van den Bosch, (2005). Memory-Based Language Processing. Cambridge University Press.

[22] Dalal A. et al. (2007). Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi. International Conference on Natural Language Processing.

[23] Dandapat S., Sarkar S. and Basu A. (2004). A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. Transactions on Engineering, Computing and Technology. V1 December 2004 ISSN 1305-5313.

[24] Dandapat S., Sarkar S. and Basu, A. (2004). Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario. Proceedings of the ACL 2007 Demo and Poster Sessions, pages 221–224, Prague, June

[25] Dermatas E. and George K. (1995). Automatic stochastic tagging of natural language texts. Computational Linguistics, 21(2): 137-163.

[26] DeRose S. J. (1988). Grammatical category disambiguation by statistical optimization. Computational Linguistics, 14:31-39.

[27] Devi S.R. (2013). Is Manipuri an Endangered Language? Language in India www.languageinindia.com ISSN 1930-2940 Vol. 13:5 May 2013.

[28] Fry J. (2007). Part-of-Speech Tagged Corpora. Linguistics 115: Corpus Linguistics, Fall 2007, SJSU.

[29] Garg N. et al. (2012). Rule Based Hindi Part of Speech Tagger. Proceedings of COLING 2012: Demonstration Papers, pages 163–174. COLING 2012, Mumbai, December 2012.

[30] Graeme D. R., Alan W. B., Graham J. R. and Stephen G. P. (1992). Computational Morphology: Practical Mechanisms for the English Lexicon. Cambridge: The MIT Press.

[31] Green, B and Rubin, G (1971). Automated Grammatical Tagging of English. Department of Linguistics, Brown University.

[32] Grierson G.A. (ed.) (1903-28). Linguistic Survey of India. Vol. III, Pt. III (reprinted 1967-68). Delhi-Varanasi: Motilal Banarsidas.

[33] Hardie A. (2004). The Computational Analysis of Morphosyntactic Categories in Urdu. PhD Thesis submitted to Lancaster University.

[34] Hajic J., Krbec P., Kveton P., Oliva K. and Petkevic V. (2001). A Case Study in Czech Tagging. In proceedings of the 39th Annual Meeting of the ACL.

[35] Halevi Y. (2006). Part of speech tagging. Seminar in Natural Language Processing and Computational Linguistics, School of Computer Science, Tel Aviv University, Israel, April.

[36] Harris Z. (1962). String analysis of the language structure. Mutton and Co., The Hauge.

[37] Heeman P.A. and Allen J.F. (1997). Incoporating POS tagging into language modeling. In Proceedings of the eight conference on European chapter of the Association for Computational Linguistics, Madrid, Spain. 230-237.

[38] HSK _ Corpus Linguistics (2008). Development of tag sets or part-of-speech tagging. MILES, Release 18.02x on Tuesday January 2218:53:50 BST, 2008.

[39] http://language.worldofcomputing.net/pos-tagging/rulebased-pos-tagging.html.

[40] http://www.ldcil.org/standardsTextPOS.aspx

[41] http://www.comp.lancs.ac.uk/computing/research/stemming/general/ paice.htm

[42] http://www.cfilt.iitb.ac.in/

[43] http://www.ethnologue.com/language/mni

[44] http://research.microsoft.com/en-us/groups/mls/

[45] IIIT-tagset. A Parts-of-Speech tagset for Indian languages. http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf.

[46] Jivani A.G. (2001). A Comparative Study of Stemming Algorithms. International Journal Computer Technology and Applications, Vol. 2 (6), pp. 1930- 1938.

[47] Jurafsky D. and Martin J. H. (2009). Speech and Language Processing (2nd ed.)”, Pearson Education (Singapore).

[48] Karthik Kumar G, Sudheer K and Avinesh Pvs. (2006). Comparative Study of Various Machine Learning Methods for Telegu Part of speech tagging. In proceedings of the NLPAI Machine Learning 2006 Competetion.

[49] Klein S. and Simmons R.F (1963). A computational approach to grammatical coding of English words. Journal of the Association for Computing Machinery, Vol. 10, pp. 334-47.

[50] Kumar D. and Josan G. S. (2010). Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. International Journal of Computer Applications (0975 – 8887) Volume6–No.5, September, 2010, www.ijcaonline.org/ volume6/number5 /pxc3871409.pdf.

[51] Lafferty J., McCallum A. and Pereira F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning. 282-289.

[52] Leech G. and Wilson A. (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R.

[53] Loftsson H. (2008). Tagging Icelandic text: A linguistic rule-based approach. Nordic Journal of Linguistics 31.1, XX-XX.

[54] Lovins J. B. (1968). Development of a stemming algorithm. Mechanical Translation and Computer Linguistic., vol.11, no.1/2, pp. 22-3.

[55] Manju K, Soumya S. and Idicula S. M. (2009). Development of a POS tagger for Malayalam – An Experience. International Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[56] Mcteer M., Schwartz R. and Weischedel R. (1991). Empirical studies in part-of-speech labeling. Proceedings of the 4th DARPA Workshop on Speech and Natural Language, pp. 331-336.

[57] Merialdo B. (1994). Tagging English Text with a Probabilistic Model. Computational Linguistics, 20(2):155-171.

[58] Megyesi B. (1998). Brill's Rule-Based Part of Speech Tagger for Hungarian. Thesis submitted to Stockholm University.

[59] Mitkov R. (2009) (2nd ed.). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press.

[60] Oflazer K. & Kuruoz I. (1994). Tagging and morphological disambiguation of Turkish text. In Proceedings of 4th ACL conference on Applied Natural Language Processing Conference.

[61] Porter M. F. (1980). An algorithm for Suffix Stripping. Program\14\ no. 3, pp 130-137.

[62] Porter M. F. revised in Nov. 2006, available at http://snowball.artarus.org/algorithms/english/ stemmer.html

[63] Rabbi I., Khan M.A. & Ali R. (2008). Developing a Tagset for Pashto Part of speech tagging. Second International Conference on Electrical Engineering 25-26 March, 2008.

[64] Ramshaw L.A. and Marcus M.P. (1995). Text chunking using transformation based learning. In Procedure Third Workshop on Very Large Corpora. ACL, 1995.

[65] Ray P. R., Harish V., Basu A. and Sarkar S., (2003). Part of speech tagging and Local Word Grouping Techniques for Natural Language Processing. In Proceedings 1st International Conference on Natural Language Processing.

[66] Samuelsson C. and Voutilainen A. (1997). Comparing a linguistic and a stochastic tagger. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (EACL), Madrid, Spain. 246-253.

[67] Santorini B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information, 1990.

[68] Shrivastav M., Melz R., Singh S., Gupta K. and Bhattacharyya P. (2006). Conditional Random Field Based POS Tagger for Hindi. In Proceedings of the MSPIL, Bombay, 63-68.

[69] Shrivastava M. and Bhattacharyya P. (2008). Hindi POS Tagger Using Naïve Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge. International Conference on NLP (ICON08), Pune, India, December, 2008. Alsobaccessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008.

[70] Singh S., Gupta K., Shrivastav M. and Bhattacharyya P. (2006). Morphological Richness Offset Resource Demand – Experience in constructing a POS Tagger for Hindi. In Proceedings of COLLING/ACL 06. 779-786.

[71] Singh T. D. and Bandyopadhyay S. (2008). Morphology Driven Manipuri POS Tagger. Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98, Hyderabad, India.

[72] Singh Ch. Y. (2000). Manipuri Grammar. Rajesh Publications, New Delhi.

[73] Singh I. (2004). Manipuri to English Dictionary. S. Ibetombi Devi, Imphal.

[74] Singh L. S., Th. K. and Das P. K. (2007). Written Manipuri (Meiteiron) – Phoneme to Grapheme Correspondence. Language in India, Volume – 7: 6 June 2007.

[75] Singh T. D. and Bandyopadhyay S. (2006). Word Class Sentence Type Identification in Manipuri Morphological Analyzer. In Proceedings of MSPIL, IIT Bombay, pp 11-17.

[76] Singha Kh. R., Purkayastha B.S., Singha Kh. D. and Roy A. (2011). Developing a Tagset for Manipuri Part of speech tagging. Journal of Computer Science and Engineering, Volume-5-issue-1-january-2011.

[77] Singha Kh. D. (2008). Loan Words in Manipuri. Bilingualism and North-East India. Assam University Publication.

[78] Singha Kh. R., Purkayastha B.S. and Singha Kh. D. (2012). Part of speech tagging in Manipuri: A rule-based Approach. International Journal of Computer Applications (0975 – 8887) Volume 51– No.14, August 2012.

[79] Singha Kh. R., Purkayastha B.S. and Singha Kh. D. (2012). Part of speech tagging in Manipuri with Hidden Markov Model. International Journal of Computer Science Issues. 9(6), 146-149.

[80] Singh J., Joshi N. and Mathur I. (2013). Development of Marathi Part of Speech Tagger Using Statistical Approach.

[81] Thoudam P.C. (2006). Problems in the Analysis of Manipuri Language. From http://www.ciil-ebooks.net, CIIL, Mysore.

[82] Wong K., Li W., Xu R. and Zhang Z. (2010). Introduction to Chinese Natural Language Processing. California: Morgan & Claypool Publishers.

Appendix A

Definitions of Terms

This appendix lists the terms frequently used in this thesis.

Computational Linguistics: Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modelling of natural language from a computational perspective.

Corpus: A corpus or text corpus is a large and structured set of texts. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

Lexical Item: A lexical item is a single word, a part of a word, or a chain of words that forms the basic elements of a language's lexicon.

Lexicon: A lexicon is a language's inventory of lexemes. The word "lexicon" derives from the Greek λεξικόν (*lexicon*), neuter of λεξικός (*lexikos*) meaning "of or for words".

Linguistics: Linguistics is the scientific study of language. There are broadly three aspects to the study, which include language form, language meaning, and language in context.

Morphosyntactic: The study of grammatical categories or linguistic units that have both morphological and syntactic properties.

Morpheme: A meaningful linguistic unit consisting of a word, such as *man,* or a word element, such as *-ed* in *walked,* that cannot be divided into smaller meaningful parts.

Morphology: A branch of linguistics that studies and describes patterns of word formation, including inflection, derivation, and compounding of a language.

Part of speech tagging: In corpus linguistics, part of speech tagging also called grammatical tagging or word-category disambiguation is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context.

Tagset: The set of tags used for annotation in a particular language in a particular corpus.

Tagged Corpus: The text corpus in which all the lexical items are annotated with its proper part of speech tag is known as tagged corpus. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistics rules within a specific language territory.

Appendix B

List of Publications

[1] Kh Raju Singha, Bipul Syam Purkayastha, Kh Dhiren Singha and Arindam Roy **“Developing a Tagset for Manipuri Part of speech tagging”** Journal of Computer Science and Engineering, Volume-5-issue-1-january-2011.
https://sites.google.com/site/jcseuk/volume-5-issue-1-january-2011

[2] Kh Raju Singha, Bipul Syam Purkayastha and Kh Dhiren Singha **“Part of speech tagging in Manipuri: A Rule-based Approach”** International Journal of Computer Applications (0975 – 8887) Volume 51– No.14, August 2012.
http://www.ijcaonline.org/archives/volume51/number14/8111-1727

[3] Kh Raju Singha, Bipul Syam Purkayastha and Kh Dhiren Singha **“Part of speech tagging in Manipuri with Hidden Markov Model”** International Journal of Computer Science Issues (1694-0814), Vol. 9, Issue 6, No 2, November 2012.
http://ijcsi.org/papers/IJCSI-9-6-2-146-149.pdf

[4] Kh Raju Singha, Ksh Krishna Bati Singha and Bipul Syam Purkayastha **“Developing a Part of Speech Tagger for Manipuri”** International Journal of Computational Linguistics and Natural Language Processing Vol 2 Issue 9 September 2013.
http://www.ijclnlp.org/vol2issue9/paper78.pdf

[5] Ksh Krishna Bati Singha, Kh Raju Singha and Bipul Syam Purkayastha **“Morphotactics of Manipuri Adjectives: A Finite-State Approach”** International Journal of Information Technology and Computer Science, 2013, 09, 94-100.
http://www.mecs-press.org/ijitcs/ijitcs-v5-n9/IJITCS-V5-N9-10.pdf

Appendix C

Participation in Conferences and workshops

Conferences

1. "5th Global WordNet Conference, 2010" organized by Indian Institute of Technology, Bombay, India from 31st January to 4th February, 2010.
2. "National Conference on Current trends in Computer Science 2010" organized by Department of Computer Science, Assam University Silchar from 22nd to 24th February 2010.
3. "National Conference on Computational Intelligence and Signal Processing CISP 2001" organized by Assam Don Bosco University, Guwahati on 2nd March 2011.
4. "National Conference on Emerging Trends and Application in Computer Science" held on 3rd March 2011 in St. Anthony's College Shillong, Meghalaya.
5. "24th International Conference on Computational Linguistics, 2012" held in IIT Bombay, India from 8th to 15th, December, 2012.
6. "1st National Conference on Research & Higher Education in Information Technology (RHEIT-2013)" held on 4th -5th February, 2013 in the Department of Information Technology, Assam University, Silchar.

Workshops

1. "National Workshop on Neural Networks and Applications" held at Gauhati University, Guwahati jointly organized by ISI Kolkata and Department of Statistics, Gauhati University from 15th to 17th December, 2009.

2. “2nd national Workshop on Indo-WordNet” held at Shillong jointly organized by Department of Computer Science, Assam University, Silchar, IIT Bombay, Gauhati University and Manipur University from 12th to 14th April, 2010.
3. “National Workshop on Language Teaching, Testing and Evaluation” organized by Department of Linguistics, Assam University Silchar from the 23rd to 26th of February, 2001 at Assam University, Silchar.
4. “National Workshop on Introduction to Natural Language Processing” from 4th to 8th March, 2013 organized by the Linguistic Data Consortium for Indian Languages, Central Institute of Indian Languages, Mysore in collaboration with the Department of Linguistics, Assam University, Silchar.