IATagger – a Tool for Tagging Indo-Aryan Texts

Krzysztof Jassem (jassem@amu.edu.pl)*

Rafał Jaworski (rjawor@amu.edu.pl)*

Krzysztof Stroński (stroniu@amu.edu.pl)*

** Adam Mickiewicz University in Poznań*

The paper aims at highlighting main functionalities of IATagger – a computer system designed to facilitate tagging early New Indo-Aryan texts.

Indo-Aryan (IA) has been a subject of corpus linguistic research for quite a long time. A number of annotated texts can be found for Old Indo-Aryan (OIA), Middle Indo-Aryan (MIA) and contemporary New Indo-Aryan (NIA). There exist well developed tagging tools for OIA and NIA (e.g. Sanskrit POStagger by Hellwig (2009) or POStagger for Urdu (Hardie 2005)).

Corpus-based research on IA has basically been carried out on OIA and MIA texts, while early NIA stages have almost been neglected. IAtagger has been designed in order to help fill this gap by optimizing the annotation of some well documented early NIA languages such as Rajasthani, Braj, Awadhi and Dakkhini.
IATagger is an environment for text annotation that provides unique capabilities whilst taking into account such issues as: productivity, flexibility and minimization of error cost.

The key functionality of the IATagger is multi-level annotation of words and sentences of early NIA texts. The default levels of word annotation are: Lexeme, Grammar (annotated using Leipzig Glossing Rules), POS, Syntax (exploring basic Dixonian (1994) scheme based on the three primitive terms: A, S and O), Semantics and Pragmatics (based on the RRG approach, e.g. Van Valin 2004). The default levels of sentence annotation are: English Translation and MetaInformation.

On request IATagger generates statistics concerning occurrences of specific classes of words and word collocations – in a specified document or collection of documents. This facilitates linguistic analysis, at the level of syntax, semantics and pragmatics. Initial analysis assumes a survey of alignment features, i.e. main argument marking with both finite and non-finite verb forms. Next steps of the analysis will also allow multifaceted analysis of IA non-finites (focusing on converbs) drawing from two frameworks: RRG (Van Valin & LaPolla 1997; Van Valin 1993; 2004) and Multivariate Analysis (Bickel 2010) where apart from morphosyntactic, semantic and pragmatic properties of converbs will be investigated.

IATagger provides several features that aim at improving the productivity of usage. For each level the system displays the context-sensitive list of prompts of available annotation tags. For a word under annotation the system displays "a prompt cloud", which consists of a set of tag suggestions for various annotation levels.
IATagger minimizes the cost of usage errors or system failure. Each annotation decision is saved automatically in a temporarily back-upped database. There is no save button.
The wide variety of configuration settings assures the flexibility of the tagger, allowing it to be used in various scenarios. It is possible to configure: the language of the annotated documents, sentence/word annotation levels as well as the annotation tags.
The system is intended for open access. It is accessible from any popular Internet browser at *http://rjawor.vm.wmi.amu.edu.pl/tagging*. The access credentials can be received on request.
We believe that IATagger capabilities, its flexibility and open access make the system attractive for use in various annotation tasks.

References

Bickel, Balthasar. 2010. *Capturing particulars and universals in clause linkage: a multivariate analysis*. In: Bril, I. (ed.) Clause Linking and Clause Hierarchy : Syntax and Pragmatics. Amsterdam: John Benjamins, 51-102.

Dixon Robert M.W., 1994. *Ergativity*, Cambridge: Cambridge University Press.

Hardie, Andrew. 2005. *Automated part-of-speech analysis of Urdu*: conceptual and technical issues. In: Yadava, Y, Bhattarai, G, Lohani, RR, Prasain, B and Parajuli, K (eds.) Contemporary issues in Nepalese linguistics. Kathmandu: Linguistic Society of Nepal.

Hellwig, Oliver. 2009. *SanskritTagger, a stochastic lexical and POS tagger for Sanskrit*. In: Huet G., Kulkarni A. and Scharf P. (eds.) Sanskrit Computational Linguistics, Lecture Notes in Computer Science Vol. 5402. 266-277.

Van Valin, Robert Jr., 1993. *A synopsis of Role and Reference Grammar*. W: Van Valin Robert Jr. (ed.) Advances in Role and Reference Grammar. Amsterdam and Philadelphia: John Benjamins, 1-164.

Van Valin, Robert J. & LaPolla Randy 1997. *Syntax*. Cambridge: Cambridge University Press.

Van Valin, Robert Jr. 2004. *Exploring the Syntax–Semantics Interface.* Cambridge: Cambridge University Press.