Automatic converb detection in Early Braj

Rafał Jaworski*, Krzysztof Stroński†

*Department of Natural Language Processing
Adam Mickiewicz University in Poznań, Poland
rjawor@amu.edu.pl

†Chair of Oriental Studies
Adam Mickiewicz University in Poznań, Poland
stroniu@amu.edu.pl

Abstract

This paper presents the work on automatic converb detection in Early Braj prose and poetry from 15-17th centuries. This is a continuation of research on non-finite verbs in early New Indo-Aryan (NIA) languages (Jaworski and Stroński, 2016). The goal of the detection mechanism is to successfully identify a plaintext word as a converb or non-converb. Such mechanism facilitates further converb description and analysis, which is of great importance in research on historical syntax of NIA. In order to develop the automatic detector, a selection of state-of-art statistical classification mechanisms was used.

Acknowledgements

This research was supported by Polish National Centre for Science Grant 013/10/M/HS2/00553.

1. Introduction

Early NIA langugaes lack digitalized, let alone annotated corpora. Therefore, in comparison to Old Indo-Aryan or Middle Indo-Aryan, early NIA period which is crucial for understanding the main mechanisms of syntatic change has been understudied. An important desideratum is thus a tool detecting selected grammatical forms and facilitating their annotation. The present study which focuses only on one early NIA variety, namely early Braj is a modest attempt to fill this gap.

2. Converb in Indo-Aryan

The category of the converb has been recognized in the typological literature in early 90's and since then widely used as a crosslinguistically valid notion defined as "a non-finite verb form whose main function is to mark adverbial subordination" (Haspelmath, 1995). In the context of IA languages this is one of the most central categories due to its role in clause linking as well as in defining of what is labelled 'the South Asian Linguistic Area', cf. (Emenau, 1956), (Masica, 1976). Therefore the converb has received considerable attention in the IA linguisctics. It has been analysed from various perspectives and at various stages of the development of IA (cf. (Dwarikesh, 1971), (Tikkanen, 1987), (Subbārāo, 2012) to name just a few). So far, however, there has been no attempt to analyse non-finites using more advanced corpus oriented quantitative and qualitative methods. The first step to such an analysis is an attmept at converb detection which in turn results in creation of possibly large set of converbal forms. In the present paper we present results of converb detection in one particular variety of early NIA, namely early Braj. The corpus consists of more than 10000 words of prose and verse composed by authors such as Indrajit of Orcha (McGregor, 1968), Hita Harivamsa [1], Vishnudas (Dvivedī, 1972) and Bhushan Tripathi (Misra, 1994) between 15-17 centuries. The selection of texts was not random. The texts belong to various genres and they seem to be quite representative for early Braj. Those texts which were available exclusively in the paper version were scanned, automatically recognized and transliterated by the HindiOCR software (Hellwig, 2015).

3. Previous research on Awadhi converbs

Elswhere (Jaworski and Stroński, 2016) we have demonstrated how converb detection can facilitate a multilayered analysis of non-emebeded structures along the lines preseneted in a multivariate analysis model (Bickel, 2010). The converb detector made it possible to extrapolate the results of the analysis performed on the mannually tagged corpus to an untagged one. We have chosen quite a homogenous corpus of early Awadhi (Jayasi's Padmāvat - a verse text from the 16th century). We studied the main argument marking, subject identity constraint (SIC) and the scope of Tense (T), Illocutionary Force (IF) and Negation (NEG) operators.

In our targeted research, we selected 236 converbal chain constructions (116 from the tagged part and 120 from the untagged one). We concluded that main argument marking follows the same rules as in the case of finite verb constructions, i.e. the A [2] marking depends exclusively on the transitivity of the main verb. What is more, unmarked A's outnumber marked ones which may result from the decay of the inflectional system leading to the disappearance of the ergative alignment in eastern varieties of IA. We have also noticed that Differential Object Marking (DOM) in early NIA is in the stage of development along the animacy and definiteness lines – in the 16th century we still find unmarked animate and definite O's (see example (1)). This conforms to the previous research on O marking in

[1] http://wp.unil.ch/eniat/2015/05/hymns-by-hita-harivaṃsa/

[2] We use here consistently two of the three Dixonian primitive terms (Dixon, 1994) i.e. A - subject of a transitive verb and O - object of a transitive verb

other branches of IA (see for example (Wallace, 1981) for Nepali).
(1)

dhāi	*suā*	*lai*
wet-nurse.NOM.F.SG	parrot.**O**.NOM.M.SG	take.CVB
mārai	*gaī*	
kill.INF.OBL	go.PPP.F.SG	

’The wet-nurse having haken the parrot went to kill it’. (JayP86.1)

SIC violation seems to have less constraints than in contemporary NIA. It is permitted not only when the subject of the converb is inanimate and the converb denotes a non-volitional act (see example (2) with an implicit animate subject and volitional action) unlike in modern NIA, cf. (Subbārāo, 2012).
(2)

sakami	*haṁkāri*	*phāṁdi*	*giyaṁ*	*melā*
power	call.CVB	noose	neck.OBL.F.SG	put.PP.M.SG

‘[birds] having called [one another] with power [loudly], the noose was put on their neck.’ (JayP72.3)

Preliminary analysis of the T-scope has brought interesting conclusions pertaining to the aspectual value of the IA converb. We assume that the fact that the IA converb is not congruent with the present tense reference gives direct evidence to its perfectivity (compare examples (3) and (4) in which T scope is conjunct or disjunct respectively). Interestinlgly we could find typological parallels to it as well.
(3)

paṁkhi-nha	*dekhi*	*saba-nhi*	*ḍara*
bird.OBL.M.PL	see.CVB	all.OBL.PL	fear.M.SG
khāvā			
eat.PPP.M.SG			

‘... birds saw all of that and got scared.’ (JayP69.2)

(4)

sidha	*ḍarahiṃ*	*nahiṃ*	*apane*	*jīvāṁ*
holyman	fear.3PL.PRS	not	self	life.OBL.M.SG
kharaga	*dekhi*	*kai*	*nāvahiṃ*	*gīvāṃ*
sword	see.CVB	CVB	bow.3Pl.PRS	neck.OBL.F.SG

‘Holly men do not fear for their own life but they bow their necks having seen a sword.’ (JayP240.3)

In Awadhi, IF and NEG scope results from the position of the markers. If the Q-word is in the right most position in the clause, then its scope is prefereably conjunct but when it preceeds the main verb and the converb is preposed, the scope remains local. If the NEG marker is in front of the postposed main verb or the converbal clause the scope of negation is local. However, if the NEG marker occurs in front of the preposed converbal clause, the scope of negation is also local but it does not extend to the adjacent converbal clause. Similar conclusions were arrived at in the preliminary research on Rajasthani (Jaworski et al., 2015).

Metric	Score
Accuracy	96.91%
Precision	56.8%
Recall	55.3%
F-score	56.0%

Table 1: Results of the baseline converb detector.

4. Automatic converb detection in Braj

The aim of an automatic converb detection mechanism is to annotate all converbs in a text. The input is a plain text written in Early Braj, which is transliterated according to ISO 15919 rules. The text is split into sentences and tokenized. The detector is run on each tokenized sentence. The input for the detection mechanism is therefore a series of plaintext, unannotated words, for which the detector is to provide yes/no answers, indicating whether they are converbs or not. This makes it a binary classification problem.

4.1. Experimental data

In order to collect the data for the experiment, a portion of Braj documents were annotated manually on word level. This task was facilitated by the *IA tagger* system (Jaworski et al., 2015), available at `http://rjawor.vm.wmi.amu.edu.pl/tagging/` (login credentials can be obtained on request to the authors). The system enables the users to provide annotations for individual words in the sentence on several annotation levels. Among the levels there are: grammatical and morpho-syntactic information, POS-tags, semantics and pragmatic information (the list of levels and tags is configurable and typically tailored to a specific annotation task). The annotation is displayed as a table, where individual words are column headers and the annotation levels are represented in rows.

In our experiment, 10 001 words were tagged on the levels: grammar (according to the Leipzig Glossing Rules) and part-of-speech. The information about a word being a converb in the annotated texts is stored on the POS level in the appropriate tag: CVB. An example of an annotated Braj sentence is shown in Picture 1.

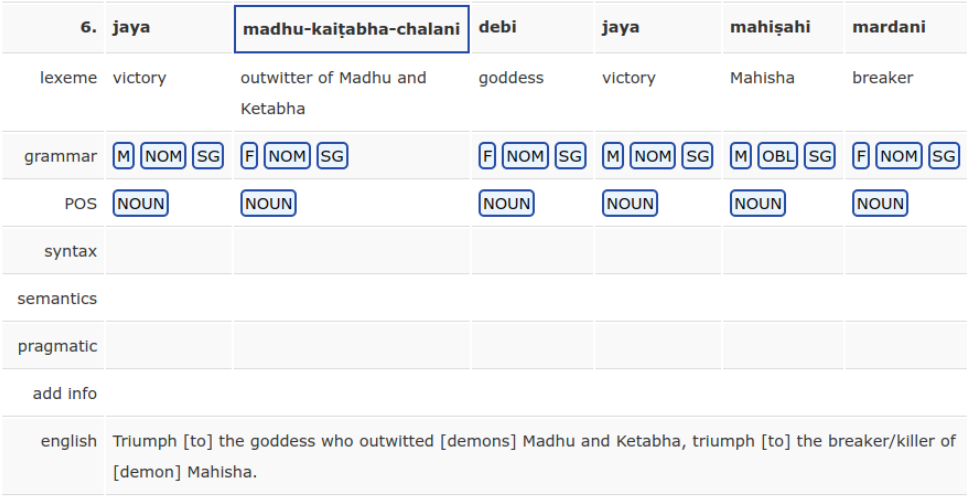
Out of the total 10 001 words, 263 were annotated as converbs, which constitutes 2.6%.

4.2. Baseline system

The first step was to construct a baseline system, which works under the following principle: all the converbs encountered in the training set are added to the dictionary. Then, all the words in the test set which appear in the converb dictionary are annotated as converbs, while the remaining words are annotated as non-converbs. This step was taken in order to establish, whether converb detection in the collected data is a trivial or non-trivial problem.

All developed binary classifiers, including the baseline, were tested in the same scenario of 10-fold cross validation. The results obtained by the baseline system are shown in Table 1.

As the converbs are rare in the training and test sets, the accuracy measure is not reliable (most classifiers un-



6.	jaya	madhu-kaiṭabha-chalani	debi	jaya	mahiṣahi	mardani
lexeme	victory	outwitter of Madhu and Ketabha	goddess	victory	Mahisha	breaker
grammar	M NOM SG	F NOM SG	F NOM SG	M NOM SG	M OBL SG	F NOM SG
POS	NOUN	NOUN	NOUN	NOUN	NOUN	NOUN
syntax						
semantics						
pragmatic						
add info						
english	Triumph [to] the goddess who outwitted [demons] Madhu and Ketabha, triumph [to] the breaker/killer of [demon] Mahisha.					

Figure 1: Braj sentence annotated in the IA tagger system

der such circumstances are heavily biased towards non-converb predictions, which are often correct). However, the precision and recall scores for the baseline system can provide useful information.

The baseline approach was supposed to yield a high precision but low recall score. The recall score was indeed low, which follows from the fact that converbs, as specific forms of verbs, are an open class and therefore there is a relatively high probability that a converb in the test set had not been detected during training.

However, the precision score achieved by the baseline system is also surprisingly low. The interpretation of this result is the following: some, not rare homonymous words can in some contexts be converbs, while in other contexts – not.

In conclusion, the converb detection problem in this scenario is definitely non-trivial.

4.3. VW and converb detection in other early NIA

Experiments in automatic processing of texts in early NIA had been done before (Jaworski and Stroński, 2017). The first phase of the research involved developing an automatic part-of-speech tagger for Rajasthani. All training data was acquired by the means of the IA tagger system. Each word was assigned exactly one POS tag, coming from a tagset of 22 tags. The part-of-speech tagging process for Rajasthani was seen by the authors as a multi-class classification problem which was approached with an algorithm based on the Maximum Entropy principle. As some of the tags in the tagset formed hierarchies (e.g. there was a "NOUN" tag and its child, "NOUN-SINGULAR") the authors computed the accuracy of the automatic POS tagger in two scenarios - the first required that the automatically assigned tag matched exactly the expected tag (exact matching), while the second allowed partial matching, e.g. assigning a NOUN tag to a word tagged as NOUN-SINGULAR. Unfortunately, the accuracy results achieved in both scenarios fell below expectations, yielding 57.9% for exact matching and 64.1% for partial matching. Separately, individual precision and recall scores for detecting only converbs were computed for this system. The precision was as low as 33% with the recall not exceeding 7%.

	Rajasthani	Awadhi
Precision	83%	80.2%
Recall	39%	64.4%
F-score	53%	71.4%

Table 2: Results of the converb detectors for Rajasthani and Awadhi.

This failure turned us towards developing systems focused on the one part-of-speech of particular interest – the converb. As a result, the authors developed specialized converb detectors for Rajasthani (Jaworski et al., 2015) and Awadhi (Jaworski and Stroński, 2017). The algorithms relied on statistical binary classifiers. The results achieved by the binary converb detectors are shown in Table 2.

4.4. Vowpal Wabbit classifier for Braj

In order to tackle the problem of automatic converb detection in Braj, the Vowpal Wabbit (VW) software (Langford et al., 2009) was used. VW is a well-established, robust statistical classification toolkit, which combines numerous classification and regression algorithms. VW is well optimized for the use of sparse features. A useful functionality of VW is the analysis of the impact of individual features on the predictions (the *vw-info* program).

The design of the converb detector for Braj was inspired by previous research on similar problems for Rajasthani and Awadhi (Jaworski and Stroński, 2017). The feature engineering process resulted in finding the following most informative features:

- three letter suffix of the word
- two letter suffix of the word
- one letter suffix of the word (i.e. the last letter of the word)
- literal form of the previous word in the sentence

Metric	Score
Accuracy	98.50%
Precision	81.8%
Recall	74.4%
F-score	77.9%

Table 3: Results of the VW converb detector for Braj.

- distributional similarity class of the previous word in the sentence
- distributional similarity class form of the next word in the sentence
- (binary feature) is the word first or last in the sentence

Distributional similarity (often abbreviated *distsim*) is a method for categorizing words in a large corpus based on their contexts. Each word falls into a category with other words that appeared in similar contexts. The id of such a category can be used as a word feature.

In order to compute distributional similarity classes, an unannotated modern Rajasthani corpus of 81 843 words was used. It was processed with the help of word2vec software, described in (Mikolov et al., 2013). The words were categorized into 209 classes, each containing between 1 and 66 words. For example, one of the classes contained the following words: *te* 'this', *teha* 's/he', *bi* 'two', *bewai* 'both', which are all pronouns.

The final results of the developed VW converb detector are shown in Table 3.

4.5. Interpretation of achieved results

The results of the VW converb detector show a considerable improvement in comparison with the baseline. They allow to use the developed converb detector in a scenario previously applied in the aforementioned research on Awadhi. It is possible to take a large unannotated collection of texts in Braj and run the converb detector on the text. Next, the sentences containing automatically detected converbs are presented to a team of linguists who manually verify the results of the detection. High precision results of the detector ensure that in most cases (4 out of 5) the linguists are presented with actual converbs, which saves vast amounts of their labour. On the other hand, the recall of nearly 75% indicates that only 1 in 4 de facto converbs in the unannotated text will not be presented to the linguists. This loss, however, is acceptable in the light of the following fact: as manual analysis of the whole large corpus would be unfeasible, such project would not be started altogether, leading to a 100% converb loss. Note also that any recall score above the score achieved by dictionary detection implies that the detector is able to detect converbs unseen in the training data. Such new converb examples are of high interest to the linguists.

5. Conclusions

In this paper we have presented results of the research on early New Indo-Aryan languages, involving traditional linguistic analysis aided by computational methods. For the purpose of annotation of early NIA corpus the IA tagger system was designed. The system allows for manual word-level text annotation on several levels, which is performed by linguists. Collected data serves both for direct linguistic analysis and for training and testing automatic classifiers.

As the recent research on early Braj is focused on specific verb forms – converbs – automatic converb detector for Braj was developed. The mechanism is a binary classifier, which relies on the Vowpal Wabbit machine learning toolkit. Features for the binary classifier are based on the literal form of a word, its context and the information on distributional similarity. The performance of the classifier allows for its use in a scenario, where converbs are automatically detected in a large unannotated text collection and then verified by linguists. This method facilitates the process of a multilayered linguistic analysis of any type of converbal constructions.

6. References

Bickel, Balthasar, 2010. Capturing particulars and universals in clause linkage: a multivariate analysis. In Isabelle Bril (ed.), *Clause Linking and Clause Hierarchy : Syntax and Pragmatics*, number 121 in Studies in Language Companion Series. Amsterdam: John Benjamins, pages 51 – 102.

Dixon, Robert M.W., 1994. *Ergativity*. Cambridge Studies in Linguistics. Cambridge University Press.

Dvivedī, Loknāth, 1972. *Viṣṇudās kavkiṛt Rāmāyana kathā*. Sāhitya bhavan limited.

Dwarikesh, Dwarika Prasad Sharma, 1971. Historical syntax of the conjunctivc participle phrase in new indo-aryan dialects of madhyadesa (midland) of northern india. University of Chicago Ph.D . dissertation.

Emenau, Murray, 1956. The sanskrit gerund: A synchronic, diachronic and typological analysis. *Language*, (32):3–16.

Haspelmath, Martin, 1995. The converb as a cross-linguistically valid category. In Martin Haspelmath and Ekkehard König (eds.), *Converbs in cross-linguistic perspective: structure and meaning of adverbial verb forms – adverbial participles, gerunds*, number 13 in Empirical approaches to language typology. Berlin: Mouton de Gruyter, pages 1 – 55.

Hellwig, Oliver, 2015. ind.senz – OCR software for Hindi, Marathi, Tamil, and Sanskrit. http://www.indsenz.com.

Jaworski, Rafał, Krzysztof Jassem, and Krzysztof Stroński, 2015. Manual and Automatic Tagging of Indo-Aryan Languages. *Human Language Technologies as a Challenge for Computer Science and Linguistics*:550–554.

Jaworski, Rafał and Krzysztof Stroński, 2016. New perspectives in annotating early new indo-aryan texts. In *Proceedings of the 32nd South Asian Languages Analysis Round Table SALA-32, Lisbon, Portugal.*

Jaworski, Rafał and Krzysztof Stroński, 2017. Recognition and multi-layered analysis of converbs in early nia. In *Proceedings of the 33rd South Asian Languages Analysis Round Table SALA-33, Poznań, Poland.*

Langford, John, Lihong Li, and Tong Zhang, 2009. Sparse online learning via truncated gradient. In *Advances in neural information processing systems*.

Masica, Colin P, 1976. *Defining a linguistic area: South Asia*. Chicago University Press.

McGregor, R.S., 1968. *The Language of Indrajit of Orchā: A Study of Early Braj Bhāsā Prose*. University of Cambridge Oriental Publications. Cambridge University Press.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Misra, Visvanath P., 1994. *Bhusana granthavali*. Nai Dilli, Vani Prakasan.

Subbārāo, K.V., 2012. *South Asian Languages: A Syntactic Typology*. South Asian Languages: A Syntactic Typology. Cambridge University Press.

Tikkanen, B., 1987. *The Sanskrit gerund: a synchronic, diachronic, and typological analysis*. Studia Orientalia. Finnish Oriental Society.

Wallace, William D, 1981. Object-marking in the history of nepali: a case of syntactic diffusion. *Studies in the Linguistic Sciences*, 11(2).