Binary Classification Algorithms for the Detection of Sparse Word Forms in New Indo-Aryan Languages

Rafał Jaworski(✉), Krzysztof Jassem, and Krzysztof Stroński

Adam Mickiewicz University in Poznań, Poznań, Poland
rjawor@amu.edu.pl

Abstract. This paper describes experiments in applying statistical classification algorithms for the detection of converbs – rare word forms found in historical texts in New Indo-Aryan languages. The digitized texts were first manually tagged with the help of a custom made tool called IA Tagger enabling semi-automatic tagging of the texts. One of the features of the system is the generation of statistical data on occurrences of words and phrases in various contexts, which helps perform historical linguistic analysis at the levels of morphosyntax, semantics and pragmatics. The experiments carried out on data annotated with the use of IA Tagger involved the training of multi-class and binary POS-classifiers.

1 Introduction

The aim of the present paper[1] is twofold. Firstly, it attempts to give a brief overview of a tool designed for semi-automatic annotation of early New Indo-Aryan (NIA) texts. Secondly, it focuses on several aspects of automatic POS-tagging of early NIA.

There has been a considerable amount of corpus-based research into Old and New Indo-Aryan, which has already contributed much to our knowledge and understanding of the history of one of the main branches of Indo-European. However, there is still a need for research into early NIA. The present paper is a modest contribution to the corpora collation and preliminary analysis of early NIA morphosyntax from various perspectives. We decided to focus on selected early NIA tongues such as Rajasthani, Awadhi, Braj, Dakkhini and Pahari, and we present here the preliminary results of research into the Rajasthani language, based on short prose texts from the 15th to 18th centuries supplemented by early Awadhi poetry[2].

[1] This paper is a part of a research project funded by Polish National Centre for Science Grant 2013/10/M/HS2/00553.

[2] Optical recognition of Rajasthani texts was supported by a Hindi OCR program [11].

Z. Vetulani et al. (Eds.): LTC 2015, LNAI 10930, pp. 123–136, 2018.
https://doi.org/10.1007/978-3-319-93782-3_10

2 The IA Tagger Tool

2.1 Tool Overview

IA Tagger is a tool for text annotation specifically for Indo-Aryan languages. The key functionality of the tool is multi-level annotation of words and sentences of early NIA texts (see Sect. 2.2). IA Tagger provides several features that improve the efficiency of use. For most annotation levels the system displays a 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 (see Sect. 2.3).

IA Tagger minimizes the cost of usage errors or system failure. Each annotation decision is saved automatically in a periodically backed-up database. There is no save button.

The wide variety of configuration settings ensures the flexibility of the tagger, allowing it to be used in various scenarios (see Sect. 2.4).

On request IA Tagger generates statistics concerning occurrences of specific classes of words and word collocations – in a specified document or collection of documents (see Sect. 2.5).

The system is intended for open access. It is accessible using any popular Internet browser at http://rjawor.vm.wmi.amu.edu.pl/tagging. Access credentials can be obtained on request from *rafal.jaworski@amu.edu.pl*.

2.2 Multi-level Tagging

Upon upload to the system, a document is automatically split into sentences (see Fig. 1).

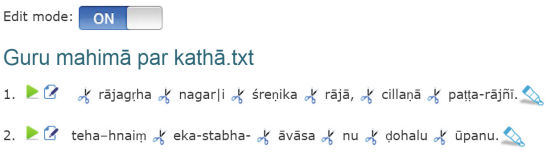

Fig. 1. Sentence split.

The user can easily override the automatic sentence split (using "scissors" or "glue"; Fig. 1). The document is annotated in sentence-by-sentence mode.

Each sentence is automatically split into words. The user may override the word split, e.g. in order to divide a word into a stem and a suffix. Words are annotated at six levels: Lexeme (where the closest English lexical equivalent is given), Grammar (annotated by means of Leipzig Glossing Rules), POS (Parts of Speech), Syntax (exploring the basic Dixonian [7] scheme based on the three primitive terms: A, S and O, where A stands for the subject of a transitive sentence, S for the subject of an intransitive sentence and O for the object of a

transitive sentence), Semantics (where we distinguish six basic thematic roles: Agent, Patient, Experiencer, Recipient, Stimulus and Theme, based on the RRG approach, e.g. [30]), and Pragmatics (distinguishing Topics).

Figure 2 represents an annotated sentence in Rajasthani.

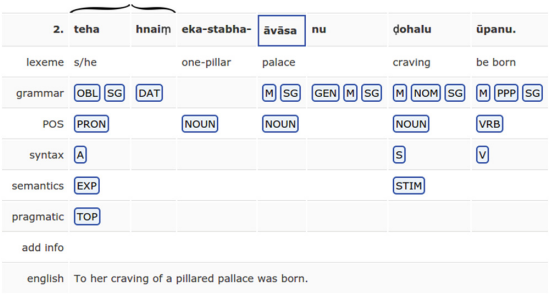

Fig. 2. Annotated sentence.

2.3 Automatically Generated Suggestions

In order to improve tagging efficiency, the system suggests hints whenever possible, i.e. when a word has already been tagged or when the tagging could be deduced automatically. Tag suggestions appear in a "cloud" above the word (Fig. 3).

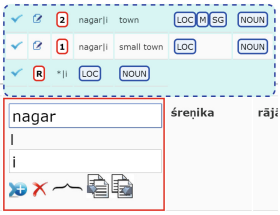

Fig. 3. Automatically generated suggestions.

Figure 3 shows tag suggestions for the word 'nagar—i' (the pipe indicates that the word 'nagari' has been split into a stem and a suffix). The first two lines come from previous annotations, whereas the third line is the set of suggestions deduced automatically. The user can accept the set of suggestions by clicking the 'check' symbol in the left-most column. The annotation shown in Fig. 2 was obtained by applying the third-line set of tags.

2.4 Configuration

IA Tagger may be configured to serve a variety of annotation tasks. The "configuration" option allows one to manage the languages of tagged documents as well as to configure annotation levels. Annotation levels may be freely ordered, added, deleted or edited. Editing of an annotation level consists in defining admissible values of respective tags.

2.5 Flexible Statistics Generator and Preliminary Results of Its Application

On request, IA Tagger generates statistics concerning occurrences of specific classes of words and word collocations – in a specified document or collection of documents. This facilitates a linguistic analysis at the levels of morphology, syntax, semantics and pragmatics. Initial analysis assumes a survey of alignment features, i.e. main argument marking (A and O). This kind of research has already been carried out by several authors for finite verbs (e.g. [13]), but IA Tagger makes it possible to generate statistics for texts belonging to various historical stages of NIA, and it has a much larger scope since it encompasses both finite and non-finite verbs such as converbs[3], infinitives and participles. Preliminary research on a Rajasthani annotated corpus shows a preponderance of unmarked O forms over marked ones. Out of 201 converbal chain constructions only 18 show marked O forms. Up to the 18th century there are only examples of animate and definite O's, or possibly human and indefinite, and from the 18th century onwards inanimate Os start appearing (see ex. (1)). Similarly, in a targeted search in an Early Awadhi text (Jāyasī's 'Padmāvat' from 1540) out of the 236 converbal chain constructions excerpted from the IA Tagger system we have found only 4 attestations of marked O's, 1 being animate and definite and 3 inanimate and definite.

(1) Old Rajasthani, 18th c. [2, 72]

tiṇa	sahanāṇa-ṇūṃ	dekha
that.OBL.SG	sign-ACC	see.CVB

'having seen that sign'

This conforms with Khokhlova's [13] findings for finites and with more general tendencies operating along definiteness and animacy hierarchies (cf. for example [1]). Morever, research on historical syntax of other early NIA tongues clearly demonstrates that O marking is rather a recent phenomenon (cf. for example [31]).

The next steps of the analysis will consist in a multilayered analysis of IA non-finites (focusing on converbs and infinitives) drawing from two frameworks: RRG [28–30] and Multivariate Analysis [3] where apart from morphological properties, syntactic semantic and pragmatic properties of converbs will be investigated.

[3] We accept here Haspelmath's [9, 3] definition of the converb: "a nonfinite verb form whose main function is to mark adverbial subordination".

Here we are going to focus briefly on control properties of converbs and scopal properties of selected operators in converbal chain constructions.

As it is the case of modern IA, converbs are predominantly controlled by the subject of the main clause (aee example (2)). However, there are single attestations of the violation of this rule. The violation of what is labelled 'the subject identity constraint' (hence SIC) has semantic or pragmatic motivation (for example, the subject of the main clause is in a possessor-like relation with the subject of the converb or it is discourse prominent - see the examples (3) and (4)). More recent, typologically oriented research by Subbarao [22] demonstrates that in IA SIC is permitted only when the subject of the converb is inanimate and the converbal clause denotes non-volitional act. As we can see from the examples (3) and (4), which has a volitional animate subject, it is not always the case of early NIA.

(2) Old Rajasthani, 15th c. [2, 13]

yakṣ-i	arjuna	ripu	bāṃdhī-karī	page	āṇi
Yaksha.INS.M.SG	Arjuna	enemy-M.NOM.SG	bind.CVB	foot	come.CVB
ghātiu					
throw.PPP.M.SG					

'Yaksha, having bound the enemy named Arjuna, threw him on his feet.'

(3) Old Rajasthani, 16th/17th c. [2, 44]

hemū [...]	pāṇīpaṃtha	āi	ḍerā	paṛiyā	chai
Hemu	Panipat	come.CVB	camp.NOM.M.PL	fall.PPP.M.PL	be.3SG.PRS

'And after that Hemu had come to Panipat, the camps were established.'

(4) Old Rajasthani, 15th c. [2, 11]

ti	puruṣa	raja-nai	vacani	karī	saṃgha-māhi
these	people	king-M.DAT.SG	speech	do.CVB	community-in
gayā					
go.PPP.M.PL					

'These men on hearing the king's speech (lit. of the king having spoken) went happy to their community.'

It seems that early NIA represents a more general South Asian type of a converb in terms of its syntactic lability - i.e. SIC violation can be permitted when we have so called *constructio ad sensum* (cf. [24]). This notion assumes a role oriented (e.g. posessor or experiencer like) relations between the converbal subject and the subject of the main clause. Actually, what we observe in (3) is a type of an absolute construction in which prototypically SIC is not observed[4]. Early NIA had absolute constructions in which the embedded clause was formed

[4] The following is a brief characteristics of an absolute constuction: "The head noun and its participle form a special type of a subordinate clause which could express an event contemporary with or anterior to that in main clause." [4].

by means of a subject modified by an inflected past participle but parallel the construction based on the converb was developed (for more detailed discussion see [25]).

Preliminary results of the analysis of scopal properties of selected operators in converbal chain constructions were presented in [21]. It has been observed that IF (Illocutionary Force) Operator can have conjunct or local scope, and this property is quite stable throughout the centuries (cf. ex. (5) and (6) from the 16 and the 17th century texts with the imperative scope local or conjunct).

(5) Old Rajasthani, 16th c. [2, 33]

paṇi	tumhẽ	mayā	karī̃	deśāntari	pahucaü
but	you	mercy.NOM.F.SG	do.CVB	abroad.OBL	reach.PPP.M.SG

'But you, having shown mercy, go abroad'.
'But you show mercy and go abroad'

(6) Old Rajasthani, 17th c. [2, 44]

upāṛi -ara	ghoṛā	māṃhe	ghālo
lift-CVB	horse.NOM.M.PL	in	throw.IMP.2PL

'After lifting the horses throw them into (the river).'
'Lift the horses and throw them into (the river).'

The T (Tense) Operator seems to have conjunct scope in those converbal chains which have the main verb in the past tense and almost exclusively local scope in those chains which have the main verb in the present tense (cf. ex. (7) and (8)). This finding has two important consequences. Firstly, it somehow implicitly presumes the perfectivity of the IA converb and secondly, it shows that converbal chain constructions are not characterised by the operator dependence. The operator dependence is a defining feature of the third type of linking, namely cosubordination which plays an important role in the theory of clause linkage in RRG [28]. Therefore, if the converbal chain constructions do not show the consistent operator dependence (and it is also the case of other operators such as IF-operator or NEG-operator) they are not instantiations of cosubordination. The support for such interpretations comes from other IA languages, both early and contemporary (see for a more detailed discussion see [23, 23-33]).

(7) Old Rajasthani, 15th c. [2, 15]

āmbā	leī̃	ḍohalu	pūriu
mango.NOM.M.PL	take.CVB	craving.M.SG	fill.PPP.M.SG

'Having taken mangos, (he) fulfilled the craving.'
'(He)took mangos and fulfilled the craving.'

(8) Old Rajasthani, 18th c. [2, 61]

phūladhārā	vica	uḍi	paṛaṃ
stream of flowers	middle.CVB	fly.CVB	fall.1PL.PRS.SBJ

'Having flown in the middle of the stream of flowers,
we shall fall'

The intermediate position of converbal chains between subordination and coordination can be investigated in terms of the operator scope which in IA may have a multiple motivation. Several authors have already tried to look at the problem from syntactic, semantic and pragmatic angles and the views are quite divergent (cf. for example [6,12]). It seems that only a more fine grained approach can bring interesting results. First attempts have already been made in synchronic and typological works by Peterson [18] and Bickel [3] and now we shall be able to extrapolate these methodologies to diachronic research.

3 Automatic POS-tagging

3.1 Similar Experiments

Experiments with automatic POS-tagging of less-resourced languages have already been conducted in recent years. This subsection briefly describes the techniques used and the outcome of two projects: an automatic tagger for Urdu, developed by [8], and Sanskrittagger by [10].

Urdu Tagger. The tagger for Urdu was developed by Andrew Hardie in 2005. The main difficulty in tagging Urdu texts identified by the author was word sense disambiguation. Two techniques were implemented in order to resolve this problem. One was based on hand-crafted rules prepared by a linguist, while the other relied on statistical analysis of manually annotated Urdu texts. The author reports the low effectiveness of the latter method, attributing it to the relatively small quantity of training data. Hence the author decided to use the tagger based on hand-crafted rules. It must be pointed out, however, that the statistical model used was HMM (Hidden Markov Models), which was considered state-of-the-art in the early 2000s, but was replaced in the following years by several other methods, such as Conditional Random Fields or Maximum Entropy.

The resulting rule-based tagger used a tagset of approximately 80 tags and achieved an accuracy of 88–90%. The author admitted that these results were lower than those of taggers for well-resourced languages, such as English. Such taggers score at least 95% accuracy. This, however, should not be considered the main flaw of this system. A more important drawback of the approach presented by Hardie is the heavy reliance on manually designed rules, which account for most of the positive results of the system. These rules were specially designed to work with Urdu, and even more specifically – with the Urdu texts that were at the author's disposal. In a different scenario the same rules may prove to be inapplicable, thus impairing the performance of the system significantly.

Sanskrittagger. Sanskrittagger, described in [10], is an automatic tokenizer and tagger for Sanskrit. Like Hardie's Urdu tagger, it uses HMM to perform the tagging. Interestingly, the same model is also applied to the task of tokenization, which is a non-standard solution.

The system uses a tagset of 136 tags. Unfortunately, accuracy figures are not known, as the evaluation of the system was performed on only five short passages of text. However, it is revealed that the system is purely statistical.

Among suggested methods of improvement, one seems particularly interesting – integrating tokenization and POS-tagging into one mechanism. The author argues that this might be a good approach for Sanskrit, even though it is not commonly used for other languages.

3.2 Training the Automated Tagger

The IA Tagger system has been used by a team of linguists for several months. The work has resulted in a manually annotated corpus of the early Rajasthani language supplemented by early Awadhi. The corpus so far contains 1284 sentences with 13 022 words. Even though the size of the corpus is too small for statistical data analysis, experiments were run to determine whether it is possible to create a usable POS tagger for early NIA.

Firstly, two separate POS tagging systems were developed. One of them uses a set of 22 tags to annotate the text. The tags are hierarchical, e.g. there is a NOUN tag and its child – NOUN-SINGULAR.

The other tagger is a detector of specific verb forms – converbs.

Multi-class POS Tagging. The task of annotation with 22 tags was seen as a multi-class classification problem. In order to implement such a tagger, a well-known Maximum Entropy tagging mechanism was used. This idea was first proposed by [19] and later used to implement the Stanford Part-Of-Speech Tagger (see [26,27]). The automatic tagger for early NIA is based on the Stanford software.

The main difficulty in training automatic taggers using the Maximum Entropy principle is the identification of the feature set. Possible features may include: suffix(n) of the word (i.e. last n letters), length of the word, whether the word starts with a capital letter (boolean feature) and many others. It is crucial, however, that all these features should be computable on unannotated text. Thus, features like "is located between a noun and a verb" are not acceptable.

The described automatic tagger for early NIA texts uses the following set of features: *Suffix(6)*, *Previous word (i.e. the literal text form of the previous word)*, *Next word* and *Distributional similarity class*.

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 [17]. 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.

Converb Detector. The second approach involved the training of a separate tagger, focused solely on identifying words of special interest – converbs. This is a case of binary classification. Two such binary converb detectors were implemented – one based on the Maximum Entropy algorithm and another one using the Vowpal Wabbit library [14].

The implementation of the converb detector using the Maximum Entropy (ME) algorithm is based on the Python NLTK library, described in [15], which is capable of using an optimization technique called MEGAM [5]. This makes it possible to create a robust binary classifier. This converb detector was trained on the same data as the multi-class tagger described in Sect. 3.2 The features used by this detector are presented in Table 1. Note that the features *cvbEnding* and *firstOrLast* use linguistic knowledge about converbs. Firstly, Rajasthani converbs typically terminate in /i/ and /a/, although from the earliest texts onwards other suffixes are also attested. Secondly, converbs would never appear as the first or last word in the sentence. This approach recalls the hand-crafted rules as seen in [8]. However, the features are never strict. The decision on whether or not to use a specific feature is made by the statistical model.

Table 1. Features used by the ME converb detector

Feature name	Parameters	Description
word	none	literal text of the word
wordContext	n	n words to the left and n words to the right of the word
suffix	n	n last characters of the word
class	none	distributional similarity class
classContext	n	as in wordContext, only on *distsim* classes
cvbEnding	none	whether or not the word ends in a typical converb ending
firstOrLast	none	whether or not the word is first or last in the sentence

In a separate experiment, a second converb detector was build with the help of the Vowpal Wabbit (VW) software and was trained on a set of 5596 Awadhi words. All these words came from texts authored by the same person. Because of the homogeneity of the texts, we expected better evaluation results than in the previous experiments.

On the other hand, the classification algorithm used by the Vowpal Wabbit software allows is based on classic regression and features numerous improvements described thoroughly in [14]. Importantly, the software features a tool for assessing the importance of individual features in the process of prediction. The most informative features identified with the help of this tool were used in the process of classification and are presented in Table 2.

Table 2. Features used by the VW converb detector

Feature name	Description
Suffix(3)	Last three letters of the word
Context(1)	The literal forms of the previous and next word
Class-context(1)	The distributional similarity classes of the previous and next word

3.3 Experiment Results

This section presents the results of the experiment conducted using both of the automatic POS taggers. In both cases the tagged corpus (13 022 words) was used to perform 10-fold cross-validation. The magnitude of the test data complies with the standards for human evaluation experiments in the field of natural language processing (see for instance [20]).

Table 3 presents results for the multi-class tagger. It assigned tags to 10 730 out of 13 022 words (82.4%), leaving the remaining words untagged. Exact tag matching counts a tag as correct only if it matches exactly the tag in the golden standard. Partial tag matching allows, for example, the tagging of a NOUN-SINGULAR with the tag NOUN.

Table 3. Overall results of the multi-class tagger

Metric	Correct tags #	Accuracy
Exact	6210	57.9%
Partial	6874	64.1%

Some specific word forms were investigated more thoroughly. Table 4 presents precision, recall and F-measure scores (as proposed in [16]) for identifying these forms. All results assume the partial tag matching metric.

Table 4. Detailed performance of the multi-class tagger

Word form	Precision	Recall	F-measure
Verb	0.61	0.70	0.65
Noun	0.41	0.52	0.46
Past participle	0.70	0.60	0.64
Converb	0.33	0.07	0.11

The accuracy of the multi-class tagger, which was as low as 64%, was not a satisfactory result. However, the results in Table 4 reveal that even though the

overall accuracy of the system is low, some word forms can be detected more accurately, such as verbs. However, converbs, the forms of our special interest, were detected poorly by the multi-class tagger. This inspired further study using the specialized converb detector.

The detector was expected to attain higher precision and recall scores in finding converbs than the multi-class tagger. The scores of the Maximum Entropy detector are presented in Table 5. These indeed show a considerable improvement over the multi-class tagger (see Table 4). This justifies the decision to implement a separate detector solely for word forms of particular interest.

Table 5. ME converb detector scores

Metric	Value
Precision	0.83
Recall	0.39
F-score	0.53

This success inspired further work on binary classification of converbs in other texts. Experiments were carried out on 5596 Awadhi words coming from a 16th century text "Padmāvat" by Malik Muhammad Jayasi. Out of all these words, 181 were annotated as converbs, which constitutes 3.2%.

First experiments with Awadhi were carried out in a 10-fold cross-validation scheme. Firstly, a baseline system was tested. The baseline built a dictionary of converbs from its training data and applied it on the test data to make the predictions. Such approach rendered the precision and recall scores of 46.7% and 57.0% respectively, which proved that the converb detection problem on the provided data is non-trivial. The best results in this scenario were achieved by the Vowpal Wabbit converb detector described above: precision of **80.2%** and recall of **64.4%**.

The VW detector trained on the whole corpus of 5596 words was then tested in another experiment. An excerpt of 11 501 words from a different part of the "Padmāvat" was tagged manually with only the converb tags and used as test data. The results of converb detection on this fragments were **74.8%** of precision and **66.4%** of recall.

Satisfactory results of the VW converb detector made it possible to use it in the following scenario. The detector was run on the whole "Padmāvat" and provided a list of words predicted as converbs. The sentences containing these words were given to human annotators for further analysis. Thus, the linguists received a set of sentences which exhibited a high probability of containing a converb. With regard to the value of the precision measure achieved in experiments, this probability can be assessed as approx. 75%. Furthermore, the value of the recall measure suggests that about two thirds of all converbs from the "Padmāvat" were successfully detected with this method. This way, the linguists obtained

a large number of example sentences containing words of their interest without the necessity of manual annotation of the whole corpus, which was not feasible.

4 Conclusions and Future Work

The paper demonstrates how IA Tagger – a semi-automatic annotating tool – can help perform multi-level historical linguistic analyses pertaining to morphosyntax, semantics and pragmatics. A flexible statistics generator facilitates distributional analysis of various converbal forms and analysis of main argument marking with finite and non-finite verbs. At the semantic level it can also support analysis of the control properties of converbs, and at the pragmatic level it clearly helps establish the scope of main clause level operators.

In the future, the IA Tagger can further support research on other non-finite verb forms such as infinitives and participles, and what is more, it can easily identify the main grammaticalization paths with respect to light verbs.

The data acquired with the use of the IA Tagger enabled research on automatic POS-tagging for the early Rajasthani language. The research consisted in two experiments carried out on a small set (13 022 words) of annotated data. The first study had the aim of creating a multi-class POS classifier trained on the available data. The second investigated the accuracy of a binary classifier devoted to a class that was recognized poorly in the first experiment. The experiments applied standard machine learning techniques, including the recently investigated idea of distributional similarity classes. The evaluation results proved that applying purely statistical methods to a small corpus of annotated data does not yield practically applicable results for multi-class recognition. The binary classifiers, however, achieved satisfactory values for both the precision and recall measures.

We conclude that applying state-of-the-art machine learning techniques to languages that lack large annotated corpora may be useful for binary classification. Using such binary classifiers for searching of potentially interesting forms in a large text collection can greatly reduce the human effort in language analysis and help to obtain important linguistic findings.

References

1. Aissen, J.: Differential object marking: iconicity vs. economy. Nat. Lang. Linguist. Theory **21**, 435–483 (2003)
2. Bhanavat, N., Kamal, L.: Rajasthani gadya: vikas aur prakash. Shriram Mehra and Company, Agra (1997–1998)
3. Bickel, B.: Capturing particulars and universals in clause linkage: a multivariate analysis. In: Bril, I. (ed.) Clause Linking and Clause Hierarchy : Syntax and Pragmatics, No. 121 in Studies in Language Companion Series, pp. 51–102. John Benjamins, Amsterdam (2010). https://doi.org/10.5167/uzh-48989
4. Bubeník, V.: A historical syntax of late Middle Indo-Aryan (Apabhramśa). Amsterdam studies in the theory and history of linguistic science: Current issues in linguistic theory. John Benjamins, Amsterdam (1998). https://books.google.pl/books?id=abJjAAAAMAAJ

5. Daumé III, H.: Notes on CG and LM-BFGS optimization of logistic regression, August 2004
6. Davison, A.: Syntactic and semantic indeterminacy resolved: a mostly pragmatic analysis for the hindi conjunctive participle. In: Peter, C. (ed.) Radical pragmatics, pp. 101–128. Academic Press, New York (1981)
7. Dixon, R.M.: Ergativity. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge (1994). https://books.google.pl/books?id=fKfSAu6v5LYC
8. Hardie, A.: Automated part-of-speech analysis of Urdu: conceptual and technical issues. Contemporary Issues in Nepalese Linguistics, pp. 48–72 (2005)
9. Haspelmath, M.: The converb as a cross-linguistically valid category. In: Haspelmath, M., König, E. (eds.) Converbs in cross-linguistic perspective: structure and meaning of adverbial verb forms - adverbial participles, gerunds, pp. 1–55. No. 13 in Empirical approaches to language typology, Mouton de Gruyter, Berlin (1995)
10. Hellwig, O.: A stochastic lexical and POS tagger for sanskrit. In: Huet, G., Kulkarni, A., Scharf, P. (eds.) ISCLS 2007-2008. LNCS (LNAI), vol. 5402, pp. 266–277. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00155-0_11
11. Hellwig, O.: ind.senz - OCR software for Hindi, Marathi, Tamil, and Sanskrit (2015). http://www.indsenz.com
12. Kachru, Y.: On the syntax, semantics and pragmatics of the conjunctive participle in Hindi-Urdu. Stud. Linguist. Sci. **11**(2), 35–49 (1981)
13. Khokhlova, L.: Ergativity attrition in the history of western new Indo-Aryan languages (Punjabi, Gujarati and Rajastahani). The Yearbook of South Asian Languages and Linguistics, pp. 159–184 (2001)
14. Langford, J., Li, L., Zhang, T.: Sparse online learning via truncated gradient. In: Advances in Neural Information Processing Systems, pp. 905–912 (2009)
15. Loper, E., Bird, S.: NLTK: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, ETMTNLP 2002, vol. 1. pp. 63–70. Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1118108.1118117
16. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, pp. 249–252 (1999)
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781
18. Peterson, J.: The Nepali converbs: a holistic approach. In: Singh, R., Dasgupta, P. (eds.) The Yearbook of South Asian Languages and Linguistics (2002), pp. 93–134. Walter de Gruyter, Berlin (2002)
19. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 16 April 1996
20. Tou, J.T.: Information systems. In: von Brauer, W. (ed.) GI 1973. LNCS, vol. 1, pp. 489–507. Springer, Heidelberg (1973). https://doi.org/10.1007/3-540-06473-7_52
21. Stroński, K., Tokaj, J.: The diachrony of cosubordination - lessons from Indo-Aryan. In: Proceedings of the 31st South Asian Languages Analysis Roundtable (SALA-31), pp. 59–62 (2015). Extended abstract. http://ucrel.lancs.ac.uk/sala-31/doc/ABSTRACTBOOK-maincontent.pdf
22. Subbārāo, K.: South Asian languages. A Syntactic Typology. Cambridge University Press, New York (2012). https://books.google.pl/books?id=ZCfiGYvpLOQC

23. Tikkanen, B.: The Sanskrit gerund: a synchronic, diachronic, and typological analysis. Studia Orientalia, Finnish Oriental Society (1987). https://books.google.pl/books?id=XTkqAQAAIAAJ
24. Tikkanen, B.: Burushaski converbs in their south and central Asian areal context. In: Haspelmath, M., König, E. (eds.) Converbs in cross-linguistic perspective: structure and meaning of adverbial verb forms - adverbial participles, gerunds. (Empirical approaches to language typology 13.), pp. 487–528. Mouton de Gruyter, Berlin (1981)
25. Tokaj, J.: A comparative study of participles, converbs and absolute constructions in Hindi and medieval Rajasthani. Lingua Posnaniensis, pp. 105–120 (2016)
26. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL, pp. 252–259 (2003)
27. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Hong Kong, pp. 63–70, October 2000
28. Van Valin, R.D., LaPolla, R.J.: Syntax: Structure, Meaning, and Function. Cambridge University Press, Cambridge (1997)
29. Van Valin, R.J.: A synopsis of role and reference grammar. Advances in Role and Reference Grammar, pp. 1–164 (1993)
30. Van Valin, R.J.: Exploring the Syntax-Semantics Interface. Cambridge University Press, Cambridge (2005)
31. Wallace, W.D.: Object-marking in the history of Nepali: a case of syntactic diffusion. Stud. Linguist. Sci. **11**(2), 107–128 (1981)