Conjunct verbs in Punjabi and across Indo-Aryan: a corpus study

Aryaman Arora
Georgetown University
aa2190@georgetown.edu

Abstract

I introduce a new Universal Dependencies corpus for Punjabi and investigate the syntactic behaviour of conjunct verbs across the Indo-Aryan family. I find evidence of conjunct component 'stickiness' from corpus data that supports the treatment of conjunct verbs as a single constituent. The work is a step towards better coverage of UD in Indo-Aryan and further investigation of comparative and historical linguistic questions.

1 Introduction

Punjabi is the language spoken in the 'land of fiver rivers', a historical area around the tributaries of the Indus river now partitioned into the Punjab administrative regions in India and Pakistan respectively. It has over 100 million native speakers. The prestige dialect of Punjabi is Majhi (lit. *middle*), associated with the cities of Lahore, Pakistan and Amristar, India.

Punjabi is an Indo-Aryan (IA) language. Indo-Aryan is unique among language families to have both immense diversity in the modern period as well as a continuously attested history of more than 3,000 years since the attestation of Vedic Sanskrit. This makes it very exciting for work on comparative and historical linguistics, and computational methods are necessary given the vast number of texts. Unfortunately, there are large gaps in availability of labelled data for this depth and breadth.

The contribution of this paper is two-fold: I design and annotate a Punjabi Universal Dependencies corpus, and using it and other existing UD corpora for Indo-Aryan languages I investigate the properties of **conjunct verbs**, which are NOUN-VERB and ADJ-VERB constructions that behave as one morphological unit.[1] Namely, I ask: does corpus data affirm that the host is syntactically different from other verbal arguments and is it actually sensible to treat ADJ and NOUN hosts as a single class, as many works do?

Genre	Doc.	Sent.	Tok.
misc	—	71	1664
news	3	71	1274
editorial	1	39	762
blog	1	33	806
Total	5	214	4506

Table 1: Data in the Punjabi UD corpus by genre. Columns are 'documents', 'sentences', and 'tokens'.

2 Designing a Punjabi corpus

For the purpose of having a broader selection of Indo-Aryan languages to examine, I created a syntactically-annotated Universal Dependencies (Nivre et al., 2016, 2020) corpus for Punjabi in the Gurmukhi script.[2] While the corpus is relatively small, it covers several genres of text (news, editorial, and blog) and is of much higher quality than existing large treebanks for Indo-Aryan languages due to being hand-annotated.

2.1 Text composition

Table 1 shows the breakdown of text in the corpus. Given the limited time for the final project, I prioritised text diversity instead of having a large corpus of a single kind of text (which would have been easier to annotator given intra-genre language conventions). I found texts on my own and vetted them manually for quality before annotation.

Why not use existing corpora? There are already several Punjabi corpora for NLP applications. The largest one is IndicCorp with 773 million tokens (Kakwani et al., 2020). For unlabelled data, Punjabi is no low-resourced language. However, after annotating a small portion of data from

[1] A terminological note: The verb component of a conjunct verb is called the *light verb*, and the other component (regardless of part of speech) is called the *host*. Conjunct verbs are a kind of *complex predicate*, the other main subtype in IA being VERB-VERB constructions.

[2] Released here.

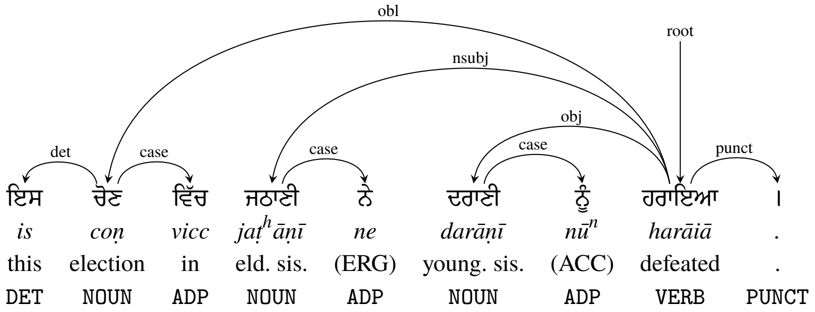

Figure 1: A Universal Dependencies-annotated sentence (id `news_bbc_inlaw_25`) from my Punjabi corpus. An English translation is "In this election, the elder sister-in-law defeated the younger sister-in-law."

IndicCorp, it became apparent that the text was low-quality, and an uncomfortably large portion of the source data could be traced back to spam websites advertising questionable products.[3] IndicCorp also tosses out document-level structure, while coherent documents could be useful to have for future multilayer annotation.

However, I did find some more carefully collected corpora. The FLORES-101 low-resource machine-translation dataset (Goyal et al., 2021), PMIndia (Haddow and Kirefu, 2020) and EMILLE (McEnery et al., 2000; Baker et al., 2002) will eventually be incorporated. I wanted more direct control over text genres though, so only small parts of FlORES have been incorporated so far.

2.2 Annotation

I annotated POS (part-of-speech) tags and dependency relations following the Universal Dependencies schema. Morphological features have not been annotated yet, but will be in a semi-automated fashion eventually. To annotate I used UD Annotatrix, a locally-hosted tool for editing `conllu` dependency trees (Tyers et al., 2017). Texts were segmented into sentences manually and tokenised by whitespace, with further manual corrections. Each document is named by its genre, source, and a unique one-word identifier, e.g. `news_bbc_rajnikanth` is a news article from the BBC about South Indian actor Rajnikanth's entry into politics.

I relied on reference dictionaries (RCPLT, 2021; Singh, 1895) and grammars (Bhatia, 1993; Gill and Gleason, 2013) to design the annotation guidelines, and also referred to other treebanks (particularly HDTB[4] and Hindi PUD[5]). The Universal Dependencies community also helped deal with some linguistic issues in annotation.[6]

Lang.	**Ref.**	**Sent.**	**Tok.**
Hindi	Tandon et al. (2016)	17.6k	375.5k
Urdu	Bhat and Sharma (2012)	5.1k	138.1k
Magahi	—	0.6k	7.7k
Bhojpuri	Ojha and Zeman (2020)	0.4k	6.7k
Punjabi	this work	0.2k	4.5k
Marathi	Ravishankar (2017)	0.5k	3.5k
Kangri	—	0.3k	2.5k
Odia	—	0.05k	0.4k
Bengali	—	0.06k	0.3k

Table 2: New Indo-Aryan UD corpora. (Sindhi UD is excluded because there it has no dependency structures.)

As a heritage speaker of Punjabi and a native speaker of the closely-related Hindi–Urdu, I also had sufficient experience with the language to be able to analyse constructions that have not been described in grammars.

2.3 Other IA corpora

Out of the New Indo-Aryan (NIA) languages, only 9 have active UD corpora with annotations, with this new Punjabi corpus being the tenth. Their sizes are listed in table 2.

3 Conjunct verbs

Conjunct verbs are an areal phenomenon of (but not exlusively of) the South Asian region, being found in both the Indo-Aryan and Dravidian families (Puttaswamy, 2018). For this corpus study,

[3] For example: `https://pa.eferrit.com/`. I am unable to understand what the purpose of these types of websites is, but all the articles felt machine-translated and unsuitable for annotation.

[4] Hindi Dependency Treebank

[5] Parallel Universal Dependencies

[6] The GitHub issues I created all dealt with copular constructions: Copula with clausal argument, What even is a copula, Copulas besides ਹੋਣਾ in Punjabi, ADJ + ਹੋਣਾ compounds in Punjabi.

two classes of conjunct verbs are under consideration, exemplified below in Punjabi:

(1) mai^{n} ne bataur ekṭar **naukrī**$_{\text{host}}$ **kītī**$_{\text{lv}}$.
I ERG as actor career did.
'I had a job as an actor.'

(2) mai^{n} ne kamre nū^{n} **sāf**$_{\text{host}}$ **kītā**$_{\text{lv}}$.
I ERG room ACC clean did.
'I cleaned the room.'

The *host* is the element providing the semantics and much of the argument structure of the conjunct verb construction, and the choice of *light verb* merely indicates transitivity and provides tense-aspect-mood information. (1) has a `NOUN` host and (2) has an `ADJ` host.

Extensive theoretical linguistic work on IA conjunct verbs (Burton-Page, 1957; Hacker, 1961; Kachru, 1982; Mohanan, 1994, 1995; Vaidya, 2015; Montaut, 2016; Fatma, 2018) has led to agreement on the following points:
1. The host does not take case marking or other modifiers (e.g. determiners in the case it is a noun).
2. The host is an argument to the verb, as evidenced by agreement, but at the same time forms a morphological unit with the verb (evidenced by limitations on movement).
3. Both the host and the light verb play a role in the argument structure of the clause, but the semantics are largely provided by the host.

This also provides an easy diagnostic for whether something is a conjunct verb construction.

3.1 Adjectival conjunct verbs

However, much of the theoretical work focuses on noun hosts to the detriment of adjectives; e.g. Mohanan (1994) assumes all discussion of noun conjuncts applies to adjectives. The following examples illustrate issues in the syntactic analysis of adjectival conjunct verbs:

(3) a. mai^{n} ne kamrā **sāf**$_{\text{host}}$ **kītā**$_{\text{lv}}$.
I ERG room clean did.
'I cleaned the room.'

b. kamrā **sāf**$_{\text{host}}$ **hoiā**$_{\text{lv}}$.
room clean became
'The room was cleaned [by someone].'

(4) kamrā sāf hai.
room clean is
'The room is clean.'

In (3), we can use different light verbs (*karnā* 'to do' and *hoṇā* 'to be') to change the transitivity of the adjectival conjunct construction. Meanwhile, (4) is just an attributive copular construction, but it uses the same verb *hoṇā* in the predicate as the intransitive conjunct verb. Also note the existence of other verbs which can behave as attributive copulae, such as *baṇnā* 'to become', *rahiṇā* 'to remain/-continue to be'.

Why then do we analyse adjectival conjunct verbs as conjuncts in the first place? Why not treat the whole class of verbs (including transitive *karnā*) as taking a predicative complement, described under Universal Dependencies as `xcomp`? I will investigate the available UD corpora to gain some more evidence about the properties of conjunct verbs.

4 Analysis

In all NIA language UD corpora, conjunct verbs use the dependency relation `compound` or its subtype `compound:lvc` (which I followed in Punjabi). To run all analyses I used Python scripts, the `conllu` package for parsing UD corpora, and `plotnine` for graphs.

4.1 Claim 1: Hosts in conjunct verbs stick

In New Indo-Aryan languages, consistuent order is discourse-configurational, i.e. it is 'free' but SOV is unmarked[7] and other orderings of constituents are conditioned by pragmatic considerations and topicalisation.

One common claim is that conjunct verb hosts are 'sticky'; they cannot move in the sentence with the same flexibility as actual semantic arguments. Mohanan (1994) categorically claims that in Hindi the host can never detach from the light verb. This is claimed to be evidence that they form a single morphological unit, since per usual syntactic tendencies in Hindi verbal arguments are free to move.

To check whether conjunct hosts are 'stickier' than direct objects (`obj`, `dobj`), I first checked how far each direct object was from its expected position immediately before the verb, ignoring conjunct hosts. Example measurements (in italics is the direct object):

(5) mai^{n} ne *kamrā* **vek^{h}iā**$_{\text{v}}$. (distance: 0)

(6) mai^{n} ne *kamrā* **sāf**$_{\text{host}}$ **kītā**$_{\text{lv}}$. (0)

(7) *kamrā* [mai^{n} ne] **sāf**$_{\text{host}}$ **kītā**$_{\text{lv}}$. (1)

[7] In Kashmiri and some other more northern IA languages, V2 word order is unmarked instead, but in our sample only SOV-unmarked languages are represented.

Treebank	**Obj**	*n*	**Host** (NOUN)	*m*	**Host** (ADJ)	*k*
Bengali	0.09	22	0.00	3	—	—
Bhojpuri	0.47	55	0.77	347	1.18	38
Hindi (HDTB)	0.35	10378	0.10	8463	0.05	4813
Hindi (PUD)	0.21	1154	0.06	224	0.02	219
Magahi	0.37	385	0.08	36	—	—
Kangri	0.40	63	0.18	57	0.08	12
Marathi	0.27	181	0.04	27	0.00	5
Odia	0.63	43	**1.17**	18	—	—
Punjabi	0.28	151	0.01	69	0.05	57
Urdu	0.38	4061	0.09	4561	0.06	2486

Table 3: Mean distance of objects (ignoring hosts) and conjunct hosts from their head verb across NIA languages. Red indicates a non-significant difference. **Bold** indicates a statistically significant different in the opposite of expected direction: objects are 'stickier'. Rest are significant for hosts being stickier at $p < 0.05$.

(8) mai^{n} ne **sāf**$_{\text{host}}$ **kītā**$_{\text{lv}}$ *kamrā*. (1)

Then I calculated the same distances for conjunct hosts. To see if there is a statistically significant difference between objects and predicative complements vs. conjunct hosts, I ran a permutation test (with 1,000 permutations) to compare mean distances.

Results are shown in table 3, with figures for only NOUN comparisons in the appendix (appendix A). In almost all Indo-Aryan languages, conjunct hosts are indeed significantly stickier than objects. In Bengali for NOUNs and Kangri for ADJs the difference is not significant, likely due to small sample size. In Odia for NOUNs the result is flipped, but again the sample size is small. However, in Bhojpuri there is both a decent sample size and non-significant difference in distance for both types of conjuncts, indicating syntactic differences from the rest of Indo-Aryan that are worth investigating. Generally though, I find this claim upheld by the data.

4.2 Claim 2: Predicative complements aren't sticky

Unfortunately, in all the treebanks the number of adjectival predicative complements (ADJ with deprel `xcomp`) was quite small. In the two largest treebanks (Hindi-HDTB and Urdu) I was able to run sensible permutation tests since there was enough data. With 320 `xcomp` to test against in Hindi and 195 in Urdu, a statistically significant greater stickiness of adjectival hosts was indeed found. The average distance of `xcomp` was close to 0, but the difference was there—perhaps `xcomp` arguments can be moved freely but due to rarity stay in the unmarked position.

This suggests there is actually something special about adjectival hosts with respect to stickiness, and my line of argumentation in §3.1 (that adjectival conjuncts might be better analysed as actual arguments) is not really supported by data, since we would expect arguments to be more mobile. So, this claim is **not supported**.

5 Limitations

A major limitation of this study is that I have not been able to test the other major property of conjunct verbs: the contribution of hosts to argument structure. I do think this is feasible with the corpus study but I fear the limited coverage of infrequent lexemes will make it harder to study with these annotated UD corpora—and I am limited in space.

Also, I have poor coverage of languages here besides Hindi–Urdu in both theoretical background and corpus data; of course, one contribution of mine is the Punjabi UD corpus which is one step towards improving breadth in UD.

6 Conclusion

Syntactically annotated corpora enable the study of many interesting questions in Indo-Aryan comparative linguistics, and they have not been adequately employed for that purpose or developed to cover the family well. This paper presents both a new UD corpus for Punjabi, a low-resourced language by NLP standards, and investigates the syntactic behaviour of conjunct verbs across Indo-Aryan languages.

I plan on expanding the Punjabi UD corpus to cover more genres (epsecially `poetry` and `social media`) and adding morphological feature annotations. I also want to expand coverage of other Indo-Aryan languages—likely next candidates are Sinhala and Sindhi.

References

Paul Baker, Andrew Hardie, Tony McEnery, Hamish Cunningham, and Rob Gaizauskas. 2002. EMILLE, a 67-million word corpus of Indic languages: Data collection, mark-up and harmonisation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. Dependency treebank of Urdu and its evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 157–165, Jeju, Republic of Korea. Association for Computational Linguistics.

Tej K. Bhatia. 1993. *Punjabi: A cognitive-descriptive grammar*. Routledge, London and New York.

John Burton-Page. 1957. Compound and conjunct verbs in Hindi. *Bulletin of the School of Oriental and African Studies*, 19(3):469–478.

Shamim Fatma. 2018. *Conjunct verbs in Hindi*, pages 217–244. De Gruyter Mouton.

Harjeet Singh Gill and Henry A. Gleason. 2013. *A Reference Grammar of Punjabi*. Punjabi University, Patiala. Revised edition by Mukhtiar Singh Gill.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation.

Paul Hacker. 1961. On the problem of a method for treating the compound and conjunct verbs in Hindi. *Bulletin of the School of Oriental and African Studies*, 24(3):484–516.

Barry Haddow and Faheem Kirefu. 2020. PMIndia – a collection of parallel corpora of languages of India.

Yamuna Kachru. 1982. Conjunct verbs in Hindi–Urdu and Persian. *South Asian Review*, 6(3):117–126.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Anthony McEnery, Paul Baker, Rob Gaizauskas, and Hamish Cunningham. 2000. EMILLE: building a corpus of South Asian languages. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*, University of Exeter, UK.

Tara Mohanan. 1994. *Argument structure in Hindi*. Center for the Study of Language (CSLI).

Tara Mohanan. 1995. Wordhood and lexicality: Noun incorporation in Hindi. *Natural Language & Linguistic Theory*, 13(1):75–134.

Annie Montaut. 2016. Noun-verb complex predicates in Hindi and the rise of non-canonical subjects. In *Approaches to Complex Predicates*, pages 142–174. Brill.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Atul Kr. Ojha and Daniel Zeman. 2020. Universal Dependency treebanks for low-resource Indian languages: The case of Bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38, Marseille, France. European Language Resources Association (ELRA).

Chaithra Puttaswamy. 2018. Complex predicates in South Asian languages: An introduction. *Journal of South Asian Languages and Linguistics*, 5(1):1–3.

Vinit Ravishankar. 2017. A Universal Dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 190–200, Prague, Czech Republic.

RCPLT. 2021. *Punjabi–English Dictionary*. Punjabi University, Patiala.

Maya Singh. 1895. *The Panjabi Dictionary*. Munshi Gulab Singh Sons, Lahore.

Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat, and Dipti Sharma. 2016. Conversion from paninian karakas to Universal Dependencies for Hindi dependency treebank. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 141–150, Berlin, Germany. Association for Computational Linguistics.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2017. UD annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17, Prague, Czech Republic.

Ashwini Vaidya. 2015. *Hindi Complex Predicates: Linguistic and Computational Approaches*. Ph.D. thesis, University of Colorado at Boulder.

A Figures

See next page.

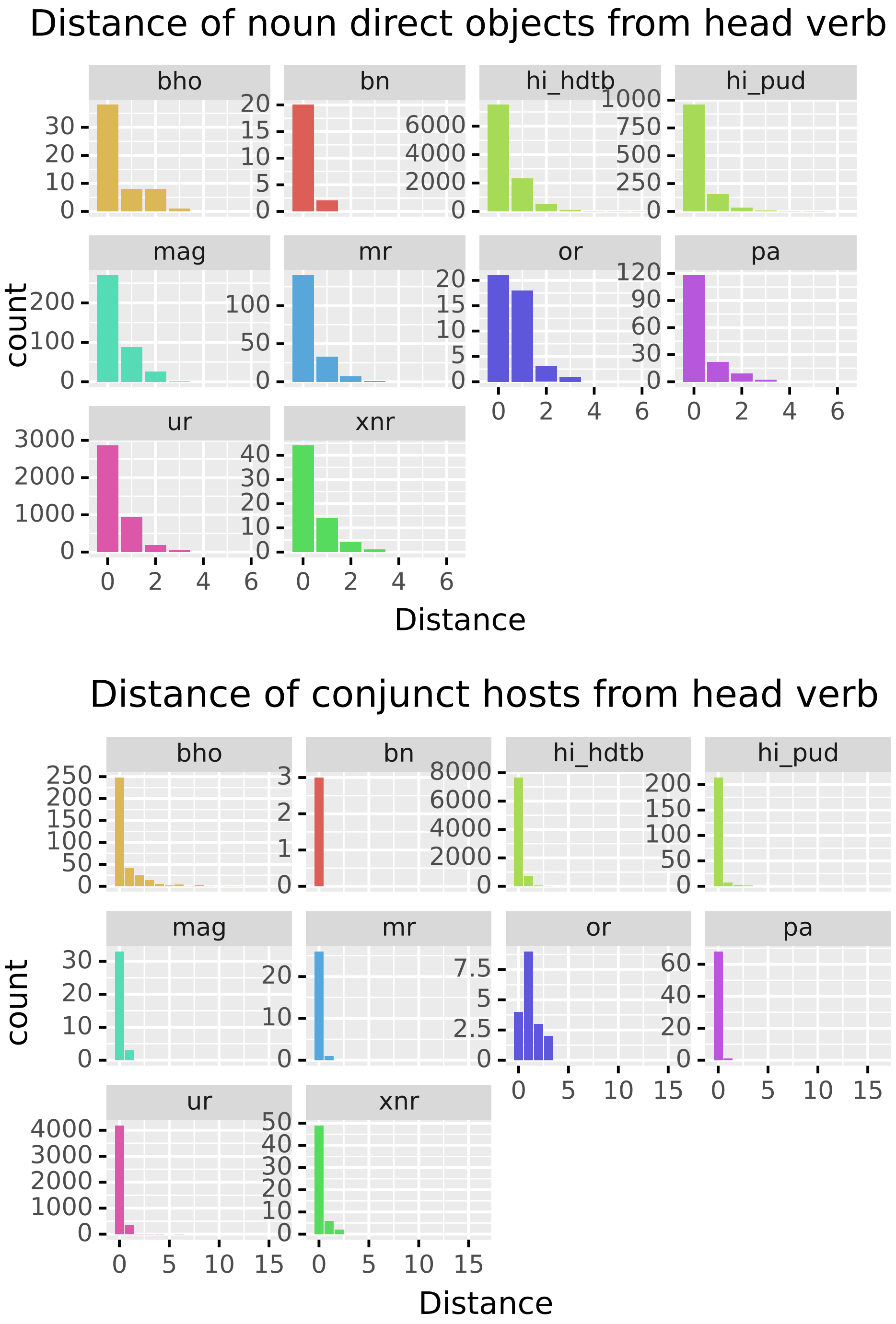

Figure 2: Distance of direct objects and conjunct hosts with POS `NOUN` from unmarked position across Indo-Aryan UD corpora.