Investigating the Probability of External Causation in Hindi Light Verb Constructions

Kanishka Jain, Ashwini Vaidya
Indian Institute of Technology, Delhi
{kanishka, avaidya}@hss.iitd.ac.in

1 Introduction

The predominant approach to analyze 'causal-noncausal' alternation in linguistics is by showing that one of the forms tends to be more coded (morphologically or phonologically) than the other (Haspelmath, 1993). For instance, in Hindi, the causal (or causative) form of the verb freeze takes the causative morpheme /-va/ (/jəm/ 'freeze' – /jəm-va/ 'caused to freeze'). Hence, the causal form is more coded than the noncausal one in this case.

Haspelmath in his series of works (Haspelmath, 2008; Haspelmath et al., 2014; Haspelmath, 2016, 2021), further extends the idea by introducing the notion of 'form-frequency' correspondence and predictability. He proposes that the form and frequency of a lexical item are correlated such that items that are more frequent are less coded or shorter compared to infrequent items. In case of the causal-noncausal alternation, if the noncausal verb is more frequent than its causal counterpart then the causal form is more coded resulting in a 'causal alternation'. But if it is the causal form that is more frequent, then the noncausal form takes an extra coding which is known as an 'anticausal alternation'. Table 1 shows examples from Swahili.

			gloss	C	NC
causal alternation	*gandisha*	***ganda***	freeze	20	**82**
anticausal alternation	***vunja***	*vunjika*	break	**883**	336

Table 1: C= causal occurrence, NC= noncausal occurrence. Verb pairs from Swahili such that in a causal alternation the noncausal form is more frequent than the causal form, and vice-a-versa in case of anticausal alternation (Haspelmath, 2008).

Causal-noncausal alternations, as in the above table, also reflect on the lexical properties of a verb such that verb pairs forming a causal alternation like 'freeze' are spontaneous events. They occur automatically without any external agent while anticausal alternations like 'break' are non-spontaneous events and occur due to the intervention of an external agent (Haspelmath et al., 2014). For instance, in English when the noncausal verb 'die' changes to the causal verb 'kill' there is an addition of external argument as shown in (1). Here, (1-a) denotes a change of state for the argument 'Sam' but (1-b) expresses the cause meaning such that John caused Sam to die. Hence, valency change is a crucial property of causal-noncausal alternations.

(1) a. Sam died.
b. John killed Sam.

However, the scope of previous studies has been limited to lexical and morphological causative alternations, and the use of other predicates as causatives have been neglected. This work aims at analyzing Light Verb Constructions (LVCs) in Hindi, where nominals alternating with the light verbs /kərna/ 'do' and /hona/ 'be' signal causal and noncausal meaning, respectively (Ahmed and Butt, 2011; Vaidya et al., 2019). Examples are shown in (2) and (3). In (2) noun /cori/ 'theft' appears with the noncausal verb /hui/ 'be' and does not require an external agent. On the other hand when the same noun appears with causal verb /ki/ 'do' as in (3) it takes an external agent /ləDka/ 'boy'. This alternation of meaning and structure is similar to our previous examples in (1).

(2) gehnõ-ki **cori** **hui**
jewellery-GEN.F theft.F be.PERF.F
'There was theft of jewellery.'

(3) lə rke-ne gehnõ-ki **cori**
boy.3.SG.M-ERG jewellery-GEN.F theft.F
ki
be.PERF.F
'The boy stole the jewellery.'

Since, light verb causal alternations as in (2) and (3) are derived from the same lexical item, that is the noun here, Haspelmath (1993) recognize them as 'equipollent' alternations or constructions with 'symmetric' coding that is both forms are coded (Haspelmath, 2021). This is in contrast with other previously investigated phenomena where one form is more coded than the other.

Further, unlike lexical and morphological causatives where the core meaning of an event comes from the verb, in case of LVCs the predicating noun carries the meaning of an event. Hence, properties like type of arguments and their semantic roles (like agent and patient) are also intricately tied to nouns instead of verbs. For instance, the noun /cori/ 'theft' in (1) and (2) is an agentive noun such that even when there is no agent in (1), there is still presupposition that there was an agent of the stealing event. In contrast, Hindi also has nouns like /izafa/ 'increase' that generally do not presuppose an external agent.

(4) ĩd^{h}ən-ki qimat mẽ izafa
fuel.M-GEN.F price.F in increase.M
hua hɛ
be.PERF.SG.M be.PRS.SG
'There is an increase in the price of fuel.'

In Hindi, the argument structure of LVCs is also dependent on the lexical properties of the nouns. For

Proceedings of the Society for Computation in Linguistics (SCiL) 2025, pages 395-397.
Eugene, Oregon, July 18-20, 2025

instance, nouns like /bɛt^{h}ək/ ‘meeting’ in (5) when occurs with the causal verb /kərna/ ‘do’ they take only one argument /məntri/ ‘minister’. While nouns like /vɪcar/ ‘thought’ in (6), when they combine with the same light verb it takes two arguments, /məntri/ ‘minister’ and /prəstav/ ‘proposal’.

(5) kəl məntrɪyõ-ne
yesterday ministers.3.PL.M-ERG
bɛt^{h}ək **ki**
meeting.3.SG.F do.PERF.F
‘The ministers held a meeting yesterday.’

(6) kəl məntrɪyõ-ne
yesterday ministers.3.PL.M-ERG
prəstav-pər **vɪcar** **kɪya**
proposal3.SG.M-on thought.3.SG.M do.PERF.M
‘The ministers considered the proposal yesterday.’

Nouns also have selection restrictions on the light verbs such that not all light verbs can combine with a noun to form an LVC (Butt, 2010). For example, nouns like /yad/ ‘memory’ can occur with different light verbs forming different LVCs (/yad kərna/ ‘intentionally remembering something/someone’, /yad hona/ ‘having a memory of someone/something’, /yad ana/ ‘unintentionally remembering something/someone’) but nouns like /pəres^{h}ani/ ‘trouble’ can only combine with light verb /hona/ (pəres^{h}ani hui ‘had a trouble’).

Considering how nouns affect both the structure and meaning of an LVC, it is interesting to ask if nouns in such constructions also affect the causalness of an LVC in Hindi. Therefore, this paper extends the notion of causality to the predicating nouns. In particular, we ask if the frequently expressed meanings can help us identify a causal or anticausal alternation for the nouns in a light verb construction. This is crucial for identifying the argument structure of the predicating noun and predicting the likelihood of the light verbs it may occur with. This also helps to build lexical resources like subcategorization frames.

2 Encoding Causalness

In this work we are interested in the general likelihood that a noun occurs more with the causal verb /kərna/ ‘do’ or with the noncausal verb /hona/ ‘be’. We show that nouns occurring more frequently with the light verb /kərna/ carry agent-oriented semantics while those that occur more frequently with /hona/ do not.

Similar to previous works (Haspelmath, 2008; Samardžić and Merlo, 2012, 2018), we study the /kərna/-/hona/ alternation by extracting their frequency distribution from syntactically annotated corpus of Hindi. We use the corpus to generate a list of alternating nouns. Following Haspelmath et al. (2014), we then calculate their degree of causalness for an LVC by dividing the total number of /kərna/ alternation multiplied by 100 by the sum of its /kərna/ and /hona/ alternations. Nouns that have high degree of causalness tends to occur more as causatives and nouns with low degree tends to occur more as inchoative. As discussed above, in /kərna/-/hona/ alternation there is no derived or marked form therefore evaluation in terms of form-frequency correspondence is not possible. Hence, to check for the validity and variability of our findings we test for agency and reproducibility. In Hindi, one way to express agency is via using ergative marker /-ne/ on the subject. We test if the agentive nouns have high probability of occurring with the ergative marker than others. We then show that this pattern is observable in other corpus of the language as well. For this, we find a list of commonly occurring LVCs across these corpora under study and found that the predicates have similar distribution.

3 Method and Results

Noun	gloss	Caus HUTB	Caus HTB	%E HUTB	%E HTB
g^{h}oʃɲa	announcement	97.7	83.3	82.9	62.5
fɛsla	decision	93.9	60	74.7	40
palən	compliance	87.5	90	37.5	0.0
ʃadi	marriage	57.1	61.1	42.8	55.6
bɛt^{h}ək	meeting	37.5	66.7	15.6	41.7
prarəmb^{h}	start	25	33.3	25	16.7
izafa	increase	16.7	28.6	8.3	0.0

Table 2: A sample of alternating LVC pairs from HUTB corpus and HTB. Caus=Causalness, %E= percentage of ergatives

To find the LVCs having /kərna/-/hona/ alternation, we have selected the Hindi-Urdu Dependency Treebank (HUTB) (∼ 4m tokens) (Bhatt et al., 2009). HUTB is a manually annotated corpus that already identifies LVCs by using the label ‘pof’ (part-of)’ and therefore LVcs can be automatically retrieved. Since, this work depends heavily on the number of LVCs that we find in the corpus we have taken only the news section ∼ 3.7m tokens) as the size of conversation data (∼ 25k tokens) is too small. We find the frequency of all the LVCs in which the nominal alternates with both the light verbs. A total of 121 alternating LVCs were found. However, to remove any chance occurrence from our analysis we remove pairs with frequency less than 1 for both the alternations giving us a list of 53 LVCs. A sample is shown in Table 2.

Based on their degree of causalness we can see that the nouns at the high end have higher probability of taking an external argument than those at the lower end. This further testified by the percentage of ergatives they occur with.

To check the validity of the realization of causalness for Hindi LVCs we try to find out whether an LVC shows a consistent behavior across different corpora or not. We conducted a comparative study by finding commonly occurring alternations in a different corpus. We compare our previous list of 53 nouns with the Hindi TimeBank’s (HTB) fictional crime part (∼ 0.2m tokens) (Goel et al., 2020). We found 25 such pairs that were common to both the corpora. We can see that nouns do show a general tendency to occur either as a causal item or as an noncausal item across the different corpora (as shown in Table 2).

In order to verify the extent to which ergativity is related to the causalness we’ve also calculated Spearman’s rank correlation coefficient. The coefficient amounts to 0.606 (level of significance = .01 (one-sided)), indicating a robust correlation between the two. However, the correlation coefficient for HTB amounts to 0.323 (level of significance = .01 (one-sided)). There are two reason for low correlation in HTB. First, ergative marker /ne/ in Hindi appears only with the subject of past perfective sentences and as a result this test didn't cover all the instances of the subject. Second, the size of the HTB corpus is smaller in comparison to HUTB.

4 Discussion

In this paper we've investigated nouns alternating with the light verbs /kərna/-/hona/ in terms of their causal property. Constructions like LVCs are distinctive as both the forms are coded therefore Haspelmath's original proposal of form-frequency correspondence and predictability of the shortness of the form does not translate to them[1]. Therefore, in this work we have extended the idea to investigate the property of 'causality' in nouns. We hypothesize that nouns have a preference towards the predicting verb which can be shown using the form-frequency correspondence. Nouns that carry more agent-oriented semantics prefer the causal verb /kərna/ while those that don't prefer the noncausal verb /hona/.

We conduct a corpus study and show that nouns in an LVC indeed have likelihood towards either the causal-noncausal formation. Nouns with high degree causalness encode agent-oriented semantics and tend to occur frequently with causal verb /kərna/ while those with lower values occur with /hona/. This is further verified by the correlation between causalness and ergativity for HUTB. We also found that similar patterns can be attested for the commonly occurring LVCs in a different corpus for Hindi.

However, there were limitations to our work. Since, Hindi has no fixed list for LVCs one may find an instance of an LVC in one corpus but not in others. Second, apart from ergativity, agency can also be tested using other parameters like animacy and volitionality of the subject. Our ongoing work focuses on testing the subject of an LVC on these various parameters. Lastly, unlike previous studies the numbers shown here are from one language only and in future work, we aim to conduct a cross-linguistic study.

References

Tafseer Ahmed and Miriam Butt. 2011. Discovering semantic classes for urdu nv complex predicates. In *Proceedings of IWCS 2011*.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for Hindi/Urdu. In *Proceedings of LAW III*, pages 186–189.

Miriam Butt. 2010. The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective*, pages 48–78.

Pranav Goel, Suhan Prabhu, Alok Debnath, Priyank Modi, and Manish Shrivastava. 2020. Hindi TimeBank: An ISO-TimeML annotated reference corpus. In *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS*, pages 13–21.

Martin Haspelmath. 1993. More on the typology of inchoative/causative verb alternations. *Causatives and transitivity*, 23:87–121.

Martin Haspelmath. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, 19(1).

Martin Haspelmath. 2016. Universals of causative and anticausative verb formation and the spontaneity scale. *Lingua Posnaniensis*, 58(2):33–63.

Martin Haspelmath. 2021. Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics*, 57(3):605–633.

Martin Haspelmath, Andreea S. Calude, Michel Spagnol, Heiko Narrog, and Elif Bamyacı. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics*, 50:587 – 625.

Tanja Samardžić and Paola Merlo. 2012. The meaning of lexical causatives in cross-linguistic variation. *Linguistic Issues in Language Technology*, 7.

Tanja Samardžić and Paola Merlo. 2018. The probability of external causation: An empirical account of crosslinguistic variation in lexical causatives. *Linguistics*, 56(5):895–938.

Ashwini Vaidya, Owen Rambow, and Martha Palmer. 2019. Syntactic composition and selectional preferences in hindi light verb constructions. *Linguistic Issues in Language Technology*, 17.

[1] A reviewer asked about the efficiency of coding and communication which LVCs seem to violate. According to Haspelmath (2021), constructions like Hindi LVCs are examples of a 'uniformly explicit' coding system where efficiency is less important than the explicit coding of meaning.