Ryan Ka Yau Lai

Beyond bidirectional association: Distinguishing light verb constructions from other conventionalised noun-verb combinations in modern Tibetan

1 Introduction

Light verb constructions (LVCs), consisting of a semantically light verb with the predicational information primarily expressed by a semantically heavy accompanying expression (usually a nominal), have long been studied in corpus linguistics within collocation analysis, which examines co-occurrence of linguistic forms in a corpus. Such analyses measure the strength of attraction between two forms with statistical association measures (AMs; Evert, 2005, 75): A pair of forms with a higher AM is likelier to be a conventionalised combination.

Traditionally, linguists focus on 'bidirectional' or 'symmetric' AMs, which quantify mutual attraction: the degree to which the two components of the construction tend to co-occur. They do not reflect whether one part of the collocation is more attracted to the other than vice versa. For example, in *wishful thinking*, *wishful* attracts *thinking* more than the other way around, but bidirectional AMs cannot represent this information. LVCs are particularly likely to exhibit asymmetric association: For example, in the English LVC *take a look*, the light verb nominal *look* likely attracts the verb *take* far more than vice versa. This may motivate the use of 'unidirectional' or 'asymmetric' measures of attraction (Michelbacher et al., 2007; Gries, 2013, 141-155; Gries, 2022, Sec. 2.1.5.), which have hitherto gained little attention.

In this paper, I explore how unidirectional measures can be added to the traditional toolkit of statistical measures for LVC detection in the context of modern Tibetan, focusing on Modern Literary Tibetan and the Central Tibetan variety spoken by the Dharamsala diaspora community. The paper's focus is not to directly propose a detection method, but to explore how the statistical distributions of light verb constructions distinguish them from other frequent noun-verb combinations, providing a clearer conceptual basis for improving feature engineering and model structure in future LVC detection work. Tibetan LVCs present challenges distinct from English

Ryan Ka Yau Lai, University of Georgia, Athens, USA

https://doi.org/10.1515/9783112229330-005

(Vincze et al., 2011). They are much more ubiquitous, including high-frequency items like the following:

(1) a. སློབ་སྦྱོང་བྱེད་
slob.sbyong byed
study do
'to study'

b. བསམ་བློ་གཏོང་
bsam.blo gtong
thought send
'to think (vol.)'

c. ཆང་ས་རྒྱག་
chang.sa rgyag
marriage strike
'to marry'

LVCs are so important in modern Tibetan that they outnumber simple verbs in spoken Lhasa Tibetan by 2:1 (Randall, 2016, 4). In these constructions, the verb is semantically light, but still encodes some semantic contrasts like volitionality and honorificity (see (4) and (5) for examples), and as Randall (2016, 122–177) shows, the nominal exhibits syntactic properties that suggest it is not as referential as the object of a 'regular' noun-verb combination, such as the inability to be modified by adjectives and demonstratives or replaced by interrogative words.

Compared to the vast resources on monomorphemic verb roots (e.g. Hackett, 2003; Hill, 2010; Hoshi, 2003), LVCs remain understudied: Information is scattered across dictionaries, with scant descriptions of argument structure and other usage information (though Bailey and Walker, 2004 includes argument structure information for a non-specialist audience). Statistical detection of LVCs from bodies of naturally occurring texts is instrumental to creating better datasets for teaching and researching Tibetan, for which the statistical distributional properties of LVCs must be better understood.

Computational work (Zhào et al., 2016) has previously detected Tibetan LVCs from corpora, employing statistical measures including entropy of surrounding tokens, measuring the diversity of contexts in which the combination appears, and pointwise mutual information (PMI), measuring bidirectional association. Yet these measures may also pick up other types of conventionalised noun-verb combinations in Tibetan. Consider the examples in (2). In (2a) the noun is semantically lighter than the verb: ངོ་ *ngo* 'face' only tells us that the target of the action has an appearance, a very general class consistent with many types of situations, whereas ཤེས་ *shes* 'know' tells us that the LVC is tells us that something is within the subject's knowledge. In (2b), the noun and verb convey similar information: both convey that the action is done for the purpose of enjoyment. In (2c), the noun and verb have distinct semantic contributions: 'alcohol' is a concrete object and 'drink' a concrete activity, but one cannot straightforwardly guess one from the other. These alternatives primarily differ from LVCs with respect to the relationship between the noun and the verb: while LVCs have a semantically specific noun carrying most of the semantics plus a semantically general verb encoding relatively little information, meaning is distributed differently in these other types of constructions. Thus, bidirectional association and context diversity are likely unable to fully capture these differences. Instead, ‘unidirectional’ measures, which examine both the verb’s relationship to the noun and vice versa instead of collapsing them into a single measure, may be more appropriate.

(2) a. ངོ་ཤེས་
ngo shes
face know
‘to recognise’

b. རྩེད་མོ་རྩེ་
rtsed.mo rtse
game play
‘play a game’

c. ཆང་འཐུང་
chang ’thung
alcohol drink
‘to drink alcohol’

To explore the possibility of using unidirectional measures, I first extract absolutive (i.e. unmarked case) noun-verb combinations from the Nanhai corpus (Schmidt, 2019), code each retrieved combination above a frequency threshold as an LVC or one of the above three categories, and calculate a number of measures of conventionalisation, including measures of association (including traditional bidirectional ones like PMI and G^2 but also unidirectional ones like ΔP and G^2rank), context entropy, and productivity. I conduct exploratory analyses of the results by plotting the measures using various graphs and performing unplanned (exploratory) hypothesis tests, focusing how these measures may potentially distinguish between different types of conventionalised N-V combinations. I also examine how similar the information encoded by different measures is.[1]

In the following, Section 2 reviews background information on LVCs, statistical association measures, and (a)symmetry in collocations. Section 3 describes the problem tackled by this paper. Section 4 describes the methodological approach for extracting potential conventionalised noun-verb combinations, and Section 5 presents the results of the study. Section 6 presents a discussion of the data.

[1] Data and code are available at https://zenodo.org/doi/10.5281/zenodo.10895756, last access 12/9/2025.

2 Background

This section reviews linguistic properties of LVCs in Modern Tibetan most relevant to the motivations and methodology of this paper (Section 2.1), statistical measures of LVC strength from in previous studies (Section 2.2), and collocational (a)symmetry and asymmetric measures (Section 2.3).

2.1 Linguistic properties of LVCs in Modern Tibetan

Tibetan syntax is generally verb-final. Most verb roots are monosyllabic with four stem forms – present, past, future, and imperative – often followed by auxiliaries and agglutinative morphology, especially in speech. Tibetan has a pragmatic case system containing, among others, an ergative གིས་=*gis* and an oblique ལ་=*la* (and their allomorphs), mostly non-obligatory (see Tournadre, 1996 for discussion on the ergative in Central Tibetan). 'Object' arguments are generally in the unmarked absolutive. As in most languages, the non-verbal element in an LVC resembles either an object (3a) or an intransitive subject (3b) even if they've lost subject/ object properties (such as the ability to be modified or replaced by an interrogative pronoun) synchronically (Lessan Pezechki and Tournadre, 2014, 9), so the LVC typically consists of a nominal with no overt case followed by an inflected verb,[2] though there are some expressions which also seem to have a semantically light verb, coupled with a semantically heavy case-marked nominal (3c; see also Randall, 2016, 71–91):

(3) a. གོམ་པ་རྒྱག་
gom.pa rgyag
walk strike
'to walk'

b. འབྲུག་སྐད་རྒྱག་
'brug.skad rgyag
thunder strike
'thunder strikes'

c. སིལ་བུར་འགྲོ་
sil.bu=r 'gro
piece=OBL go
'to fall into pieces'

The term 'light verb' is not standard in work on Tibetan and related varieties. LVC-like constructions, consisting of a highly conventionalised combination of a nominal (or adjective) plus a verb acting as a lexical unit, are also known as 'compound

[2] Sometimes, there is an additional absolutive argument, so the light verb nominal may be better described as caseless, not absolutive (Simon, 2011, 61).

verbs/ verbal compounds', 'phrasal verbs' (Agha, 1993, Denwood, 1999, etc.), 'complex predicates' (Randall, 2016), 'conjunct verbs' (Matthew & Sumi, 2005 as cited in Randall, 2016) or 'verbalised constructions' (Bartee, 2007), and the verbal component may be known as 'verbalisers' (Tournadre and Dorje, 1998), 'support verbs' (Simon, 2011), or 'secondary verbs' (see Randall, 2016, 8–12 for a terminological overview). Additionally, since disyllabic nominals with a monosyllabic verb is the most common combination, 'trisyllabic verb' is also common (Zhào et al., 2016).[3] These traditional terms cover constructions not typically considered LVCs in general linguistics. The term 'light verb' is also used by some authors (Zemp, 2018, e.g.) for semantically light auxiliary or serial verbs like ཚར་ *tshar* 'finish' or ཡོང་ *yong* 'come' which have distinct syntactic properties: They generally follow a tensed verb root rather than a deverbal nominal, and some exhibit properties like restricted inflectional possibilities and phonological reduction (see DeLancey, 1991 for examples from Central Tibetan).

The three most common Tibetan light verbs are བྱེད་ *byed*, གཏོང་ *gtong*, and རྒྱག་ *rgyag*, typically encoding volitional action verbs, as in (1). Some examples with non-volitional light verbs are as follows:

(4) a. ཁོང་ཁྲོ་ལང་
khong.khro lang
anger arise
'to get angry'

b. མགོ་སྐོར་ཐེབས་
mgo.skor thebs
fraud arrive
'to be cheated'

c. གད་མོ་ཤོར་
gad.mo shor
laughter let.loose
'to let out laughter'

Light verbs also encode honorific/ humilific status, as in the following:

(5) a. དགའ་བསུ་ཞུ་
dga'.bsu zhu
welcome do.HUM
'to welcome (hum.)'

b. ངོ་སྤྲོད་གནང་
ngo.sprod gnang
introduction do.HON
'to introduce (hon.)'

c. འདྲ་པར་སྐྱོན་
'dra.par skyon
photograph do.HON
'to photograph (hon.)'

[3] The terminological diversity reflects uncertainty as to whether the verb and nominal constitute one 'word'; although terms like 'verbaliser' and 'trisyllabic verb' strongly imply that they are a single word, there are reasons to consider them separate words, e.g. the fact that the noun and verb may be separated by forms like adverbs (discussed below). I will be agnostic on the issue.

There is no straightforward mapping from plain to honorific/ humilific forms. Though *skyon* is generally used for *rgyag* 'strike', *zhu* and *gnang* are far more flexible; for example, while the non-honorific version of (3) would use *byed*, *zhu* can also map to ཤོད་ *shod* 'say'.

The non-verbal element accompanying a light verb is most commonly, but not always, a deverbal nominal (e.g. *'dra.par* in (5c) is not deverbal). Moreover, most deverbal light-verb nominals are composed of two verbal roots, which do not necessarily appear frequently as main verbs in the modern language (e.g. *skad* in (3c) is uncommon as a verb). This is unlike English, where most light-verb nominals are derived from productive verbs, e.g. *look* in *take a look*. Thus, measures based on properties of the verbalised nominal in English corpus linguistics (e.g. using properties of the verb look when investigating *take a look*) cannot be straightforwardly applied to modern Tibetan.

As in English, a single nominal can take on multiple light verbs with different semantic nuances, and where the nominal is deverbal, the original verb(s) is sometimes also usable as a verb alone (6). Here, (6a) and (6b) differ in volitionality, and (6c) perhaps in aspect from (6a).

(6) a. བསམ་བློ་གཏོང་
bsam.blo gtong
thought send
'to think' (= 1b)

b. བསམ་བློ་འཁོར་
bsam.blo 'khor
thought turn
'to think of'

c. བསམ་
bsam
think
'to think'

Light verbs can also accompany adjectives. These constructions are harder to detect, since adjectives can also modify the verb – and in such cases, they may in turn have a light verb nominal before them. Light verbs can also accompany light noun-adjective combinations, e.g. སེམས་པ་བཟང་པོ་བྱེད་ *sems.pa bzang.po byed* 'have a good heart'. I thus leave the complex topic of adjective-light verb combinations to future research. Similarly, I put aside constructions where an honorific light verb accompanies a plain verb.

Syntactically, Tibetan light verbs are separable from their accompanying non-verbal element by verb-modifying forms, including adverbs, adjectives, the indefinite marker *gcig*, and the similative demonstratives *'di.'dra*/ *de.'dra* (Randall, 2016, 47–63). In this paper, I call these forms interveners. Arguments cannot intervene between the light verb and the non-verbal element. Unlike in English, accompanying nominals are usually not modifiable (Randall, 2016, 43–71), so determiner-based light verb detection methods in English (Stevenson et al., 2004; Tu and Roth, 2011) will also fail for Tibetan.

2.2 Statistical properties of LVCs across languages

In this section, I first explain some notational conventions, then use them to introduce statistical measures for LVC detection proposed in the literature, focusing only on measures applicable to Tibetan.

Notation and definitions. Measures of statistical properties of LVCs, and collocation more generally, are generally based on the following contingency table 1, adapted to the case of light verbs.

Tab. 1: Contingency table for computing statistical measures of word association.

	n	$\neg n$	Totals
v	$f(n,v)$	$f(\neg n,v)$	$f(v)$
$\neg v$	$f(n,\neg v)$	$f(\neg n,\neg v)$	$f(\neg v)$
Totals	$f(n)$	$f(\neg n)$	N

Here, f refers to frequency, n stands for nominals, v stands for verbs, ¬ means 'not', and N is the total sample of candidate noun-verb combinations currently considered. For example, if n is *look* and v is *take*, then $f(n,v)$ is the frequency of *take* + *look*, $f(\neg n,v)$ is the frequency of *take* + nouns other than *look*, and $f(\neg n,\neg v)$ refers to noun-verb combinations with neither *take* nor *look*. The bottom row and rightmost column represent column and row totals respectively: $f(n)$ refers to the frequency of *look* in noun-verb combinations overall, $f(\neg v)$ refers to the frequency of verbs other than *take*, etc.

In addition to the above frequencies, I denote estimated probabilities of a noun-verb combination belonging to a particular cell in the table, estimated using frequencies from the table, as follows:

$$p(n,v) = f(n,v)/N, p(n) = f(n)/N, etc.$$

Here, p stands for probability. (Technically, these are not actual probabilities, but relative frequencies used to estimate probabilities). Thus, in our case of *take a look*, $p(n,v)$ is the estimated probability that the verb is *take* and the noun is *look*, calculated by dividing the frequency of *take* + *look* by the total number of noun-verb combinations considered.

A final piece of notation is ‘conditional probability’. It measures the probability that a nominal will be used given the verb, or vice versa:

$$p(n|v) = f(n,v)/f(v), p(\neg v|n) = f(\neg v,n)/f(n), etc.$$

Here, the pipe | means ‘given’, so $p(n\,|\,v)$ in our example means the probability that the noun is *look* given that the verb is *take*. This is calculated by dividing the number of times *take* + *look* appeared by the number of times *take* appeared.

Measures of association strength. $f(n,v)$ itself has been used as a statistical correlate of LVC status (Tan et al., 2006, 51). Since a common assumption in the corpus-linguistic literature is that light verbs are conventionalised combinations, they can be expected to occur more frequently than more incidental noun-verb combinations.

Perhaps the most common measure for detecting LVCs is pointwise mutual information (PMI) (Stevenson et al., 2004; Tan et al., 2006):

$$PMI = \log_2\left(\frac{p(n,v)}{p(n)p(v)}\right)$$

Thus, PMI considers the how likely the noun and verb are to co-occur, normalised by how often the noun and verb occur individually. Because the verb and nominal components of LVCs are generally strongly associated with each other, previous studies have frequently used high PMI as an indication that something is an LVC. Zhào et al. (2015, 2016) the PMI approach to Tibetan.

Significance measures. Significance measures are based on significant tests that quantify how unlikely the data would be if the two forms were not associated with each other, such as χ^2 and log-likelihood ratio (G^2, Dunning, 1994). This paper will adopt G^2, given its attractive properties over χ^2 especially for skewed samples (Dunning, 1994). G^2 is based on the difference between observed and expected counts. Expected counts are the number that we would, on average, expect to see in a cell of the contingency table if there were no association between the noun and the verb. It is calculated as follows:

$$E(n,v) = Np(n)p(v), E(\neg n,\neg v) = Np(\neg n)p(\neg v), etc.$$

The G^2 statistic quantifies how much the actual table deviates from these expected values. Technically, it is the log-likelihood ratio between the null hypothesis of independence and the alternative hypothesis of dependence between the noun and the verb. It takes the log of the observed divided by the expected frequency in each cell of the contingency table, weighted by twice the observed frequency:

$$G^2 = 2\left[f(n,v)\log\frac{f(n,v)}{E(n,v)} + \cdots + f(\neg n,\neg v)\log\frac{f(\neg n,\neg v)}{E(\neg n,\neg v)}\right]$$

Though, to my knowledge, significance measures have never been applied directly to LVCs, they are widespread in collocation analysis.

Contextual measures. Apart from the association measures based on contingency table values, Zhào et al.'s (2015; 2016) studies on Tibetan also considered the entropies of the previous and next tokens. These values quantify how diverse the words preceding and following the LVC are, respectively. A more conventionalised LVC would be expected to appear in a wider range of contexts:

$$H_{prev}(n,v) = -\sum_{w_{prev}} \left(p(w_{prev}|n,v)\log_2 p(w_{prev}|n,v\right)$$

$$H_{next}(n,v) = -\sum_{w_{next}} \left(p(w_{next}|n,v)\log_2 p(w_{next}|n,v\right)$$

Here, w_{prev} and w_{next} represent distinct previous and next word types, and $\sum$ is the summation symbol. Entropy is obtained by calculating the probability of each next word type multiplied by the log of itself, and summing the results. A noun-verb combination appearing in predictable contexts has low context entropy, and one appearing in diverse contexts has high context entropy.

2.3 (A)symmetry and direction in collocation analysis

A popular notion of asymmetry or directionality in English corpus linguistics comes from Kjellmer (2014, 112–115). Kjellmer distinguished between 'left-predictive' and 'right-predictive' collocations. In left-predictive collocations like *wishful thinking*, the first form strongly predicts the second one but not vice versa; in right-predictive collocations like *from afar*, the second element strongly predicts the first one. To enhance clarity and use terminology that applies to languages with lexicalised N-V combinations regardless of word order (and writing direction), I will refer instead to 'noun-to-verb (N2V)-predicting' (where the noun strongly predicts the verb but not vice versa) vs 'verb-to-noun (V2N)-predicting' (where the verb strongly predicts the noun but not vice versa) noun-verb combinations. Thus, to use English examples, *take* + *nap* is N2V, while *spite* + *face* is V2N. Both notions will be considered when investigating statistical properties in this paper. Generally, we expect LVCs to be N2V-predicting, but not V2N-predicting.

There is relatively little work on asymmetric measures for LVC detection. One exception, inspired by Dras and Johnson's (1996) related work, is from Tan et al. (2006, 50), who propose the following measure:

$$DJ = f(n,v)f(v)$$

Thus, this measure also considers how often the verb appears alone; for two sequences with each co-occurrence frequency between the noun and the verb, the LVC with the more frequent verb would have a higher DJ value. In a similar vein, Nagy and Vincze (2011, 4–6) and Vincze et al. (2011, 118–120) see improvements in some performance measures after adding a binary feature on whether a verb belongs to the 15 most common verbs in the corpus.

Other than these purely frequency-based measures, there are two other types of asymmetric measures. Firstly, in asymmetric/ unidirectional association measures, there would be separate measures for the nominal's attraction to the verb (which would be higher for V2N-predicting combinations) and the verb's attraction to the nominal (which would be higher for N2V-predicting combinations).

Michelbacher et al. (2007, 368–369) propose two asymmetric association measures. The first is conditional probabilities, i.e. $p(v\,|\,n)$ and $p(n\,|\,v)$. The second method is rank-based: For each word, they list its collocates in terms of χ^2-statistics, and use the rank of the collocate in the list as a measure of how much the collocate is attracted to the word. The higher the value of the ranking (i.e. the lower the ranking), the less attracted the collocate is to the node. Michelbacher et al. (2011, 254–257) generalise this measure to apply to arbitrary bidirectional measures, including G^2; this paper adopts this measure, denoting the rank statistic as rank(G^2).

Gries (2022, Sec. 2.1.2) offers unidirectional association measures based on the normalised Kullback-Leibler Divergence (KLD), which measures how different a distribution is from a baseline distribution. When examining how much a verb is attracted to a nominal, we compare the distribution of verbs given the noun to the distribution of verbs in the corpus overall using the following KLD formula:

$$KLD_{v\to n} = p(v\mid n)\log_2\frac{p(v\mid n)}{p(v)} + p(\neg v\mid n)\log_2\frac{p(\neg v\mid n)}{p(\neg v)}$$

If there is no association between the verb and the nominal, there would be little difference between $p(v\,|\,n)$ and $p(v)$ and between $p(\neg v\,|\,n)$ and $p(\neg v)$, pushing the KLD close to 0. KLD is normalised with the following formula so that it lies between 0 and 1:

$$KLD^{norm}_{v\to n} = 1 - e^{-KLD_{v\to n}}$$

The nominal's attraction to the verb is calculated identically, just swapping the v and n around:

$$KLD_{v\to n} = p(n\mid v)\log_2\frac{p(n\mid v)}{p(n)} + p(\neg n\mid v)\log_2\frac{p(\neg n\mid v)}{p(\neg n)},$$
$$KLD^{norm}_{v\to n} = 1 - e^{-KLD_{v\to n}}$$

Finally, ΔP (Gries, 2013, 143–155; Desagulier, 2016, 189–195) takes the difference between the probability of the verb given the noun and the probability of the verb given all other nouns (for measuring the attraction of the verb to the noun), and the difference between the probability of the noun given the verb and the probability of the noun given other verbs (for measuring the attraction of the noun to the verb).

$$\Delta P_{n|v} = p(n \mid v) - p(n \mid \neg v), \Delta P_{v|n} = p(v \mid n) - p(v|\neg n)$$

The second type of measures examines how flexible one slot in a collocation is, given another slot, which I refer to as 'slot productivity measures'. For LVCs, this means how flexible is the verb given that we know the nominal, and vice versa? Gries (2022, Sec. 2.1.3- 2.1.4), for example, proposes two such measures. Firstly, he offers 'type frequency' (TF). In the LVC case, this means the number of distinct nouns that can appear with the verb ($TF(v)$) and vice versa ($TF(n)$). Secondly, he offers 'normalised entropy', which measures how predictable the distribution of nominals is given the verb, and vice versa:

$$H_{norm}(v \mid n) = \frac{-\sum_v p(v \mid n) \log_2 p(v \mid n)}{\log_2 f(n)},$$
$$H_{norm}(n \mid v) = \frac{-\sum_n p(n \mid v) \log_2 p(n \mid v)}{\log_2 f(v)}$$

Thus, if a verb is hard to guess from the accompanying nominal, then $H_{norm}(v|n)$ would be high; if the nominal is easy to guess from the accompanying verb, $H_{norm}(n|v)$ would be low. A summary of measures is given in Table 2.

3 Problem and predicted patterns

In this section, I will first introduce N-V combinations other than LVCs in Tibetan, along with general expectations about distributional properties from an intuitive, rather than quantitative, perspective. I then make predictions about how these constructions differ distributionally from LVCs with respect to particular measures.

3.1 N-V combinations other than LVCs in Modern Tibetan

As mentioned in Section 2.1, the term 'light verb construction' is uncommon in Tibetan linguistics. Some common alternatives like 'compound verb' or 'complex predicate' do not invoke a semantically lighter verb – perhaps precisely because

Tab. 2: List of measures reviewed in Section 2.

Measure	Symbol	Symmetry	Concept operationalised
Raw co-occurrence frequency	$f(n,v)$	Symmetric	Co-occurrence frequency
Pointwise mutual information G-squared statistic	PMI G^2	Symmetric Symmetric	Association strength
Context entropy	$H_{\text{prev}}(n,v)$ $H_{\text{next}}(n,v)$	Symmetric	Context diversity
Dras-Johnson measure	DJ	Asymmetric	Verb frequency, co-occurrence frequency
Conditional probability	$p(n\|v)$, $p(v\|n)$	Asymmetric	Association strength
G-squared rank	$\text{rank}_{(v\rightarrow n)}\ (G^2)$	Asymmetric	
	$\text{rank}_{(n\rightarrow v)}\ (G^2)$		
Normalised Kullback-Leibler divergence	$KLD_{(v\rightarrow n)}{}^{\text{norm}}$ $KLD_{(n\rightarrow v)}{}^{\text{norm}}$	Asymmetric	
ΔP	$\Delta P_{(v\|n)}$, $\Delta P_{(n\|v)}$	Asymmetric	
Type frequency	$TF(n)$, $TF(v)$	Asymmetric	Flexibility of one slot given the other
Normalised entropy	$H_{\text{norm}}(v\|n)$ $H_{\text{norm}}(n\|v)$	Asymmetric	

there are constructions placed in this traditional category where the verb is not actually lighter.

Firstly, sometimes the nominal is highly predictable from the verb, rather than vice versa, such as (2a). Another example is as follows – Randall (2016, 37) considers all these complex predicates:

(7) a. གྲོད་ཁོག་རྒྱགས་
grod.khog rgyags
stomach full
'to be full'

b. གྲོད་ཁོག་ལྟོགས་
grod.khog ltogs
stomach hungry
'to be hungry'

c. གྲོད་ཁོག་བཤལ་
grod.khog bshal
stomach cleanse
'to have diarrhea'

The verb in each case, especially (7a & 7b), is strongly associated with stomachs semantically (in (7c), *bshal* seems to be most commonly used in the 'diarrhea' sense). I call these 'light noun constructions' (abbreviated LN). The nominals generally do not seem as 'light' as light verbs: They are predictable from the more specific verbs but still carry significant semantic weight on their own. Thus, these constructions are likely highly V2N-predictive but also somewhat N2V-predictive.

Secondly, sometimes the verb and noun provide similar semantic information.

(8) a. རྐུན་མ་རྐུ་
rkun.ma rku
thief steal
'to steal'

b. ཟས་ཟ་
zas za
food eat
'to eat'

c. སེམས་ཐག་གཅོད་
sems.thag gcod
resolve cut
'to make up one's mind'

When the verb and noun are etymologically related, as in (8a) & (8b), this is called 'lexical reduplication' by Randall (2016, 113–115), or a 'cognate object construction' in other languages. While the two components are not exactly equivalent in (8c), *gcod* 'cut' has several metaphorical meanings like 'decide, solve', and hence also has relatively large overlap with the nominal. To include examples like (8c) and exclude examples like རྒྱག་གཏམ་རྒྱག་ *rgyag.gtam rgyag* 'to make harsh remarks' (Randall, 2016, 114), where the nominal contains a root in the verb but is not predictable from the verb, I call these 'mutually predictable constructions' (abbreviated Mutual), which are most likely both N2V-predictive and V2N-predictive.

Finally, there exist conventionalised collocations in Tibetan where the verb and noun have clearly distinct semantic contributions, and yet remain highly associated. While some work has put these constructions in the same category as LVCs and the two other categories above, most previous work that do explicitly delineate the boundaries of complex predicates, phrasal verbs, etc. (e.g. Agha, 1993; Bartee, 2007; Randall, 2016), have focused on distinguishing them from constructions like these, which I call Distinct. Compared to the other construction types, I expected these to be less N2V-predictive and V2N-predictive (though of course, I still expect there to be some predictive power).

(9) a. ཡི་གེ་གཏོང་
yi.ge gtong
letter send
'to send a letter'

b. དེབ་ཀློགས་
deb klogs
book read
'to read a book'

c. མིང་འདོགས་
ming 'dogs
name hang
'to give a nickname'

3.2 Metrics examined and predictions

In this study, I will test most of the measures discussed in Section 2 on their ability to distinguish LVCs from other constructions. Since I am artificially restricting discussion to a subset of the full range of frequency values (i.e. taking the most frequent ones), measures based only on raw frequency will be dropped. For G^2 ranks, I use ascending ranks to keep the intuition that higher value means higher association. Ranks are defined to be maximal when G^2 is undefined, a phenomenon caused by the noun appearing only with the verb or vice versa, indicating very high attraction. Since distributions with many low probabilities are often difficult to visualise raw, the conditional probability measure is replaced by negative log-probabilities, equivalent to 'surprisal' in information theory.

For bidirectional association and context entropy measures used in previous studies, there is no predicted difference between LVCs and the other three categories, since they primarily differ from LVCs in terms of the differential roles of the noun and verbs.

Since it is generally easy to guess what kind of LVCs will follow a nominal, but very difficult to predict in the other direction, I predict LVCs to have the highest conditional surprisal for nouns given the verb (i.e. lowest V2N-predictability) and lowest surprisal for verbs given the noun (i.e. highest N2V-predictability), except perhaps for MUTUAL constructions, where the verb is also highly predictable from the noun. By the same token, I predict LVCs to have higher G^2 ranks (since there are many nouns for each given verb) and lower KLD and ΔP for nouns given verbs, and lower G^2 ranks and higher KLD and ΔP for verbs given nouns.

For slot flexibility measures, I predict that LVCs will have the highest type frequency and normalised entropy for verbs, since they tend to be highly flexible and compatible with a large number of event-denoting nouns, but make no predictions for nouns.

Table 3 shows the measures examined in this study, along with my predictions.

4 Methodology

Zhào et al.'s approach detects light verbs based on a large proprietary dataset of raw texts. Their error analyses reveal a fair number of false positives due to incorrect tokenisation, such as inappropriately selecting the last two syllables of longer light verb nominals, and examination of their results suggests that they may have been biased towards political content.

Tab. 3: Table of expectations about measures explored in this paper.

Measure type	Measure	Symbol	Compared to LVCs ...		
			LN	**Distinct**	**Mutual**
Bidirectional association	Pointwise mutual information	PMI	No expected difference		
	G-squared statistic	G^2	No expected difference		
Context diversity	Context entropy	$\mathrm{H}_{\mathrm{prev}}(n,v)$ $\mathrm{H}_{\mathrm{next}}(n,v)$	No expected difference		
Unidirectional association	Conditional surprisal	$-\log_2 p(n \mid v)$	Lower	Lower	Lower
		$-\log_2 p(v \mid n)$	Higher	Higher	/
	Chi-squared rank	$\mathrm{rank}_{(n\to v)}(G^2)$	Lower	Lower	Lower
		$\mathrm{rank}_{(v\to n)}(G^2)$	Higher	Higher	/
	Normalised KLD	$\mathrm{KLD}_{(n\to v)}{}^{\mathrm{norm}}$	Higher	Higher	Higher
		$\mathrm{KLD}_{(v\to n)}{}^{\mathrm{norm}}$	Lower	Lower	/
	ΔP	$\Delta P_{(n \mid v)}$	Higher	Higher	Higher
		$\Delta \mathrm{P}_{(\mathrm{v} \mid \mathrm{n})}$	Lower	Lower	/
Slot flexibility	Type frequency	$TF(v)$	Lower	Lower	Lower
		$TF(n)$	No expected difference		
	Normalised entropy	$\mathrm{H}_{\mathrm{norm}}(n \mid v)$	Lower	Lower	Lower
		$\mathrm{H}_{\mathrm{norm}}(v \mid n)$	No expected difference		

Unfortunately, large corpora of modern Tibetan are not currently publicly available. Moreover, because of advances in computational resources for Tibetan computational linguistics since then – including open-source tools like *botok* (Esukhia, 2023) for tokenisation – it may be more promising to detect light verbs from a tokenised corpus, minimising false positives due to inappropriate tokenisation. I thus base my study instead on a smaller, open-source corpus (~1.3M word tokens) which is balanced and tokenised, the Nanhai Corpus (Schmidt, 2019), then process it using a variety of lexical resources.

My first step is to extract noun-verb combinations, where the noun takes no case or relator noun, and acts as an argument or argument-like complex predicate nominal (i.e. not part of an adverbial phrase modifying the verb – to use an English analogy, if one were to say *I took a walk in the park yesterday*, *took* would be paired with *walk*, not *in the park* or *yesterday*). The calculations were then based on these combinations. Thus my methodology bears similarities with covarying collexeme analysis (Stefanowitsch and Gries, 2005), though it is unclear that all the combinations extracted indeed belong to a single construction (see Section 6 for theoretical discussion).

The steps I followed were as follows. The final steps as described here are the result of many iterations of trial and error, manually noting sources of false positives and false negatives and modifying the automatic process to fix them using filters and exceptions:

- Step 1: Get candidate verbs by stemming and identifying all verbs in the corpus.
- Step 2: Get candidate nominals using existing dictionary resources.
- Step 3: For each candidate verb, if there is a candidate noun before it, and there is no intervening argument, case marker, copula or quotative marker, then classify that noun-verb combination as a candidate noun-verb combination.
- Step 4: Lemmatise the verbs and create a frequency list of candidate noun-verb combinations.

Since large POS-tagged corpora of modern Tibetan are not, at the time of writing, publicly available (though this may soon change), I used a simple rule-based method for the preprocessing steps, using dictionary resources to identify potential lexicalised noun-verb combinations.

For Step 1, I began by creating a stemmed version of the corpus by removing verbal agglutinative morphology tokenised from all words, including subordinators དུས་ *dus*, ནས་ *nas*, ན་ *na*, བཞིན་ *bzhin*, nominalisers and light nouns ཡ་ *ya*, བ *ba*, པ་ *pa*, མཁན་ *mkhan*, རྒྱུ་ *rgyu*, སྟངས་ *stangs*, སྲོལ་ *srol*, ཐབས་ *thabs*, བཞིན་ *bzhin*, ས་ *sa*, ཡུལ་ *yul*, རྩིས་ *rtsis*, auxiliary འདོད་ *'dod*, and negators མ་ *ma* and མི་ *mi*. Exceptions were made for common false positives (e.g. མ་འོངས་པ་ *ma.ongs.pa* which usually means ‘future’, not ‘not having come’). I then retrieved the POS of the stemmed (but unlemmatised) word from the Monlam dictionary (Monlam, 2016), since its JSON version contains POS information, and also determined whether it was considered a verb in Hill (2010). A word satisfying both criteria, not homonymous with a case form (other than ན་ *na*, more commonly a verb ‘to be ill’), the general extender སོགས་ *sogs* or the indefinite article forms ཞིག་ *zhig*/ ཤིག་ *shig* was considered a candidate verb.

For Step 2, I first identified words that were most likely nouns, i.e. words that either had ‘noun’ as their sole POS listed in the Monlam dictionary, or had NOUN as their most common POS tag in ACTib corpus of Classical Tibetan (Hill and Meelen, 2017). (Both resources were important, since the ACTib list lacks modern words, and Monlam is relatively small.) Some exceptions were again filtered out, and false negatives common in the Nanhai corpus were brought back into the list. These false negatives were found by searching headwords in Steinert (2023), a resource combining multiple existing lexicographic sources. For each headword, I looked for other headwords consisting of the current headword plus a verb identified in Step 1 (possibly followed by the infinitive པ་ *pa*/ བ་ *ba*). Headwords that satisfied this but were not previously identified as nouns were examined and, if I identified them as nouns, added back to the noun list.

For Step 3, I separated each text into smaller text segments with the Tibetan punctuation mark *shad* །, and within each segment, considered all the candidate verbs. For each candidate verb, I identified an accompanying caseless noun as follows: If there is a noun occurring between the verb and the previous verb (or the start of the text segment if there is no previous verb), and there are no case markers, quotative markers, copulas, or words found in neither Monlam nor ACTib intervening between the noun and the candidate verb, then I extracted the noun + candidate verb combination. A small handful of nouns, e.g. time and measure words usually appearing in adverbials, were skipped over.

Of the noun-verb combinations extracted, I excluded the ones where the noun was a relational noun; these nouns act like postpositions to the more substantive nouns they follow, and thus do not form a single unit with the light verb. I then manually went through examples where the noun and verb are not adjacent. Interveners that appeared more than twice in the corpus were examined and the entries with interveners indicative of problems were removed – this mostly included nouns not recognised by the above method (including relator nouns), phrases indicating that the noun was not directly dependent on the verb (such as subordinators), and adjectives commonly participating in LVCs.

For Step 4, for each noun-verb combination, I used Hill (2010) to determine the alternative forms of each verb stem. I then take the most common form of the verb stem in the corpus as the lemmatised form.

I went through the resulting candidate N-V combinations and marked entries appearing more than 10 times as light verb, light noun, mutually predictive or distinct, resulting in n = 155 noun-verb combinations whose statistical properties will be examined below. If a false positive was found, I examined other LVCs with the same noun or verb to determine whether they were false positives too.

5 Results

In this section, I will examine the results of each statistical measure from Section 3. For association and context diversity measures, I compare the measures across the four categories of noun-verb constructions using Wilcoxo's rank-sum tests, with three contrasts per measure: between LVCs and each of LN, Distinct and Mutual. Significance levels were at .05 with Holm-Bonferroni corrections among the three contrasts. Undefined values (caused by zero counts in the case of G^2) were removed. For the slot flexibility measures, since each value is associated with multiple N-V combinations with the same noun or verb, I instead fit a Dirichlet regression with the location-scale parametrisation, the proportion of each N-V combination type as the response variable, and LVC as the baseline category, and examined the *p*-values for whether the fixed effects of the slot flexibility measures are significantly different from 0, with the same correction. A significant coefficient for a non-LVC category indicates that the odds of a construction belonging to that non-LVC category over the LVC category varies according to the slot flexibility measure in question.

5.1 Bidirectional measures of association

Examining the PMIs of different collocations reveals that high-frequency LVCs (Figure 1), while having highly positive PMI values, actually tend to have lower PMI values than high-frequency LN, DISTINCT and MUTUAL constructions. In particular, LN and LV constructions have barely any overlap in distribution. For G^2 values, LVCs are similar to Distinct constructions, but still much lower than LN and Mutual constructions. Wilcoxon tests show that PMIs' distribution is significantly different than the other three types except for Distinct (p = 0.00280 for LN, 0.37182 for Distinct, 0.00030 for Mutual), though not G^2 values (p = 0.0467 for LN, 0.0189 for Distinct, 0.4013 for Mutual). These results heavily suggest that traditional LVC detection methods that use bidirectional association will, if anything, end up favouring types of frequent N-V combinations other than LVCs.

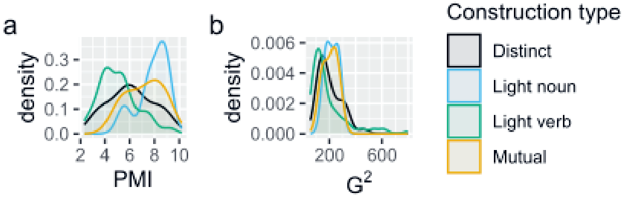

Fig. 1: Kernel density estimates with Gaussian kernels with standard deviation bandwidths for (a) PMIs and (b) χ^2.

5.2 Context diversity measures

Context entropy, as discussed by Zhào et al. (2016, 140–141), does not seem to distinguish between the different types clearly, as there is considerable overlap in the data from all the construction types, as shown in Figure 2. Wilcoxon tests show a significant difference between LVCs and DISTINCT constructions, but not the other two (preceding tokens: p = 0.1653 for LN, 0.0045 for DISTINCT, 0.9247 for MUTUAL; following tokens: p = 0.55 for LN, 0.000014 for DISTINCT, 0.90 for MUTUAL), most likely simply because DISTINCT constructions are more common, leading to a larger sample size.

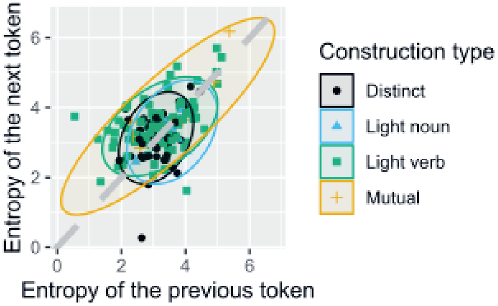

Fig. 2: Context entropy with estimated 95% bivariate normal ellipses.

5.3 Unidirectional association measures

The unidirectional association measures show very clear separation between LVCs and other types of constructions (see Figure 3). In fact, all four show the same pattern: In LVCs, the verb is attracted to the noun far more than the noun to the verb. Again, LNCs are especially divergent from LVCs.

Other than rank(χ^2), the unidirectional association measures encode virtually identical information, as shown by the scatterplots and correlations in Figure 4 on page 137.

This pattern can be easily explained. The $p(n \mid \neg v)$ value (in ΔP) and $\log_2 \frac{p(n \mid v)}{p(n)}$ value (in the KLD) are virtually negligible in this case, where a very large number of nouns corresponds to a very small number of verbs. For ΔP, this means the value is virtually equivalent to $p(n \mid v)$. For KLD, since each noun is very rare in the corpus, this means $p(\neg n) \approx 1$. Hence,

$$KLD^{norm}_{n \to v} \approx 1 - \exp(-p(\neg n \mid v) \log_2 p(\neg n \mid v))$$

which depends only on $p(\neg n \mid v) = 1 - p(n \mid v)$. Wilcoxon tests were thus excluded for ΔP and KLD. Remaining tests, shown in Table 4, found support for all differences except the ones for the verb's attraction to the noun in the cases of MUTUAL and LN.

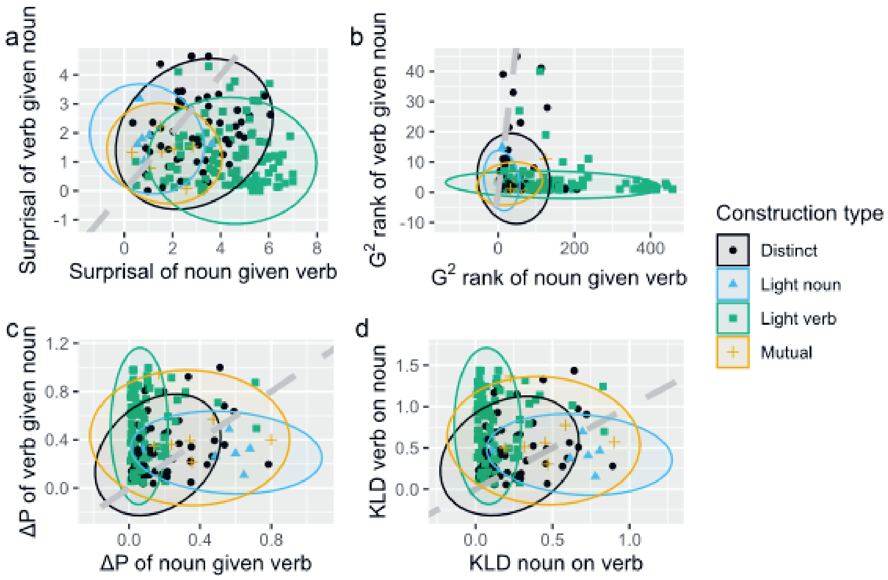

Fig. 3: (a) Surprisal, (b) G^2 ranks (with undefined values defined to have highest rank), (c) ΔP and (d) KLDs with estimated 95% bivariate normal ellipses. The dashed line indicates positions where the unidirectional measure is the same in both directions.

5.4 Slot flexibility measures

There were also clear relationships between slot flexibility measures and type of expression. Figure 5 shows the normalised entropy and type frequency values of verbs, along with a corresponding barplot showing the proportion of each category. LVCs are most common at higher values, while LNCs and to some extent Mutual are more common at the lower end, and Distinct constructions appear throughout. The opposite situation is shown in Figure 6, where LVCs are concentrated at the low end of the spectrum, and the other three categories are concentrated on the high end. Figure 7 shows that the type frequency and entropy measures are also correlated, albeit not as much as between the unidirectional association measures.

The Dirichlet regression model revealed that type frequency of nouns and normalised entropy of verbs given nouns are significantly associated with the odds of getting LNCs over LVCs and Distinct constructions, but clearly not for Mutual ones (Table 5). Likely due to the low power, there was no comparable result for verbs.

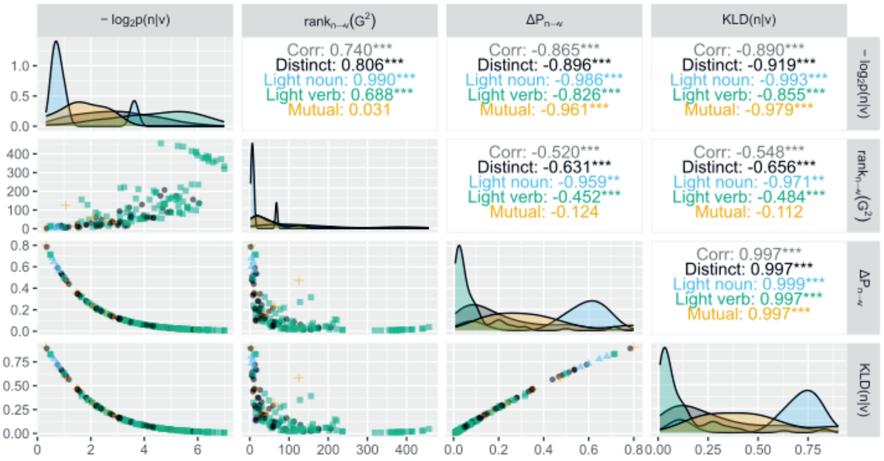

Fig. 4: Correlations between surprisal, G^2 ranks, ΔP and KLDs, for measurements of the noun's attraction to the verb. Surprisal, ΔP and KLDs are virtually just rescalings of each other. A similar pattern holds for the verb's attraction to the noun but is not shown here.

Tab. 4: *P*-values of Wilcoxon tests between LVCs and the three other types, with significant results (at .05 significance level and Bonferroni correction with g = 3) shaded.

LVCs …	with LN	with DISTINCT	with MUTUAL
$-\log_2 p(n\|v)$	0.00040	6.0×10^{-7}	6.3×10^{-5}
$-\log_2 p(v\|n)$	0.081	1.4×10^{-7}	0.44
$\mathrm{rank}_{(n\rightarrow v)}(G^2)$	0.00024	1.3×10^{-11}	0.00023
$\mathrm{rank}_{(v\rightarrow n)}(G^2)$	0.2623	0.0016	0.5598

5.5 Combining measures

Summing up the results above, we get Table 6. Differences that are significant but small are written as 'slightly', and differences that are insignificant but visually clear and have relatively small *p*-values are written as 'possibly' higher/ lower.

A question that arises from these results is whether the various measures that show differences between LVCs and the three other construction types provide overlapping or redundant information. A principal components analysis of the above measures (other than the context diversity measures, replacing G^2 with χ^2, which has the same patterns, because of the presence of undefined values for G^2) finds that 87.5% of variation in the data can be represented in three dimensions. The variables are plotted in Figure 8 and the individuals in Figure 9.

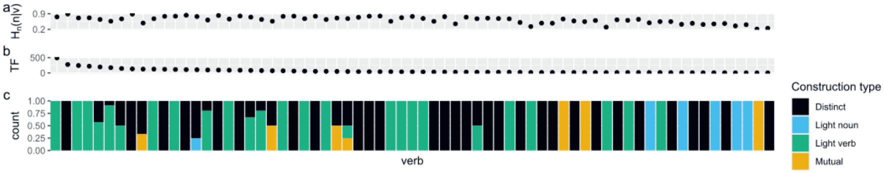

Fig. 5: (a) Normalised entropies of the noun given the verb, (b) type frequencies for each verb, and (c) proportion of the four constructions for each verb.

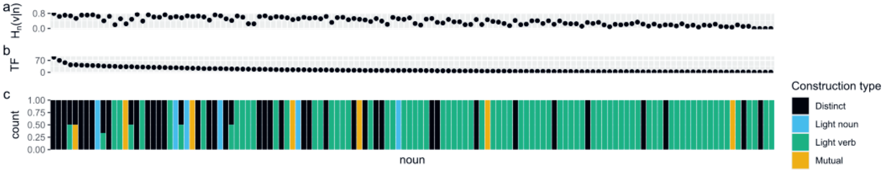

Fig. 6: (a) Normalised entropies of the verb given the noun, (b) type frequencies for each verb, and (c) proportion of the four constructions for each verb.

The first dimension (42.6% of the variance) is primarily correlated positively with high PMI and χ^2 (i.e. bidirectional association) and negatively with measures associated with lighter verbs/heavier nouns (i.e. verbs are more attracted to the noun, and nouns are more flexible given the verb); thus, LVCs tend to have very low values and LNCs high values, with MUTUAL/DISTINCT in between. The second dimension (29.9%) is mostly associated with high bidirectional association and low values of measures associated with lighter nouns/heavier verbs; hence the MUTUAL and LN constructions, where nouns are more predictable from the verb tend to have the highest values. The third dimension (8.8%) is dominated by high noun entropy, and weakly separates LNCs and MUTUAL (lower values) from some of the LVCs and DISTINCT constructions (higher values). The large overlaps in arrows in the diagrams indicate considerable overlap in the information encoded by different measures, with three main ‘clusters’ of measures indicating light nouns/heavy verbs, heavy verbs/light nouns and bidirectional association respectively, with the entropy of the noun given the verb being independent from all of these.

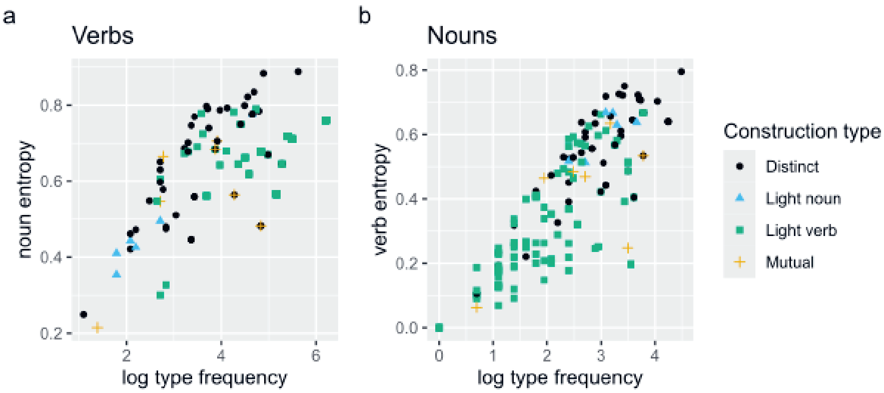

Fig. 7: Scatterplots of the relationship between logged type frequency and entropy for verbs (a) and nouns (b).

Tab. 5: *P*-values of Wald tests for the significance of Dirichlet regression coefficients, with significant results (at .05 significance level and Bonferroni correction with *g* = 3) shaded.

LVCs ...	with LN	with DISTINCT	with MUTUAL
TF(*v*)	.133	.166	.0502
TF(*n*)	2.19×10^{-7}	.0147	.956
$H_{\text{norm}}(n\|v)$	.158	.868	.697
$H_{\text{norm}}(v\|n)$	$1.82{\times}10^{-12}$	.00427	.314

6 Discussion

6.1 Asymmetry in measuring and detecting co-occurrence

The results of this paper show that, while linguists have traditionally assumed that LVCs can be detected using high bidirectional association, LVCs actually have weaker bidirectional association than other conventionalised N-V pairings at high frequencies. Thus, the corpus linguist looking to detect LVCs may need to consider a nonlinear relationship between bidirectional association and LVC status to tease LVCs apart from other types of constructions.

More generally, collocation and covarying collexeme analysis would benefit from looking beyond bidirectional measures of association, and start combining them with asymmetric measures, including both unidirectional association measures and slot flexibility measures (Gries, 2013, 2022). In this case I found substantial redundancy between different measures – including some variables that, because

Tab. 6: How different measures distinguish LVCs from other types of conventionalised noun-verb combinations. Significant results are shaded in grey.

Measures type	Measure	Compared to LVs …		
		LN	DISTINCT	MUTUAL
Bidirectional association	PMI	Much higher	/	Higher
	G^2	Possibly higher	/	Possibly higher
Context diversity	$H_{\text{prev}}(n,v)$	/	Slightly higher	/
	$H_{\text{next}}(n,v)$	/	Slightly lower	/
Unidirectional association	$-\log_2 p(n\mid v)$	Lower	Lower	Lower
	$-\log_2 p(v\mid n)$	Possibly higher	Higher	/
	$\text{rank}_{(n\rightarrow v)}(G^2)$	Lower	Lower	Lower
	$\text{rank}_{(v\rightarrow n)}(G^2)$	Possibly higher	Higher	/
Slot flexibility	$TF(v)$	Possibly lower	Possibly lower	Possibly lower
	$TF(n)$	Higher	Higher	/
	$\text{H}_{\text{norm}}(n\mid v)$	Possibly lower	/	/
	$\text{H}_{\text{norm}}(v\mid n)$	Higher	Higher	/

of the properties of the current dataset, are virtually rescalings of each other, but there remain at least three major clusters of variables, plus an additional variable (noun entropy) independent of the rest; they all seem to play different roles in distinguishing the four types of constructions.

6.2 Slot asymmetry and the construction hierarchy

As mentioned in section 2, Tibetan linguistics literature typically uses not the term LVC, but other terms like ‘compound verb’ and ‘complex predicate’ that consist mostly of LVCs, but also contain constructions of the other three categories. For example, Randall’s (2016) complex predicate includes many LN and MUTUAL constructions, and even at least one DISTINCT construction ཁོག་པ་སྔོག་ *khog.pa sngog* ‘inside’ + ‘dig’ = ‘ferret out information’.

While LVCs and LNCs both typically have one relatively productive and one relatively unproductive slot, DISTINCT constructions have two productive slots and MUTUAL ones are lexicalised with two unproductive slots. Borrowing terminology from the literature on serial verb constructions (a different type of complex predicate; Aikhenvald and Dixon, 2006), the first two are asymmetric and the second two are symmetric. Randall’s grammatical tests show that there are constructions from the four categories which can all be seen as belonging to one overarching complex predicate construction, with LVCs being particularly asymmetric and hence

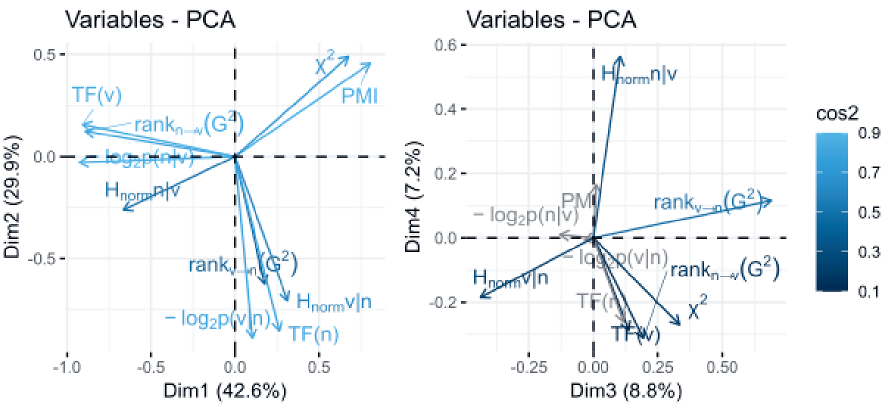

Fig. 8: Map of the different variables on the first four principal dimensions. The colours indicate squared cosine values: The larger the number, the greater the amount of variation inside that variable is represented by the dimensions in the graph.

the fixed verbal slot (light verbs) acts like a grammatical verbaliser, and the rest having less grammaticalised properties. This situation resembles what Lai and Pang (2023) describe for the Cantonese causative-resultative construction: this serial verb superconstruction also subsumes asymmetric subconstructions with the first slot highly productive and second one unproductive or vice versa, and relatively symmetric subconstructions where both slots are highly (un)productive, contra established assumptions that serial verb constructions in a specific language can be enumerated and classified simply into symmetric vs asymmetric. The Tibetan data suggests that a similar situation obtains for complex predicates derived from the dereferentialisation of the argument in argument-verb constructions.

7 Conclusion

In this paper, I explore several measures of conventionalisation as they apply to high-frequency noun-verb combinations in Tibetan to see how well they distinguish light verb constructions from other constructions, including light noun constructions, mutually predictive constructions, and constructions whose components have distinctive semantics. I find context diversity does not clearly distinguish between them. Bidirectional association measures, whose high values are typically taken to be strong indicators of LVC status, turn out to be lower for LVCs than the other three. By contrast, unidirectional association and slot flexibility measures better distinguish between LVCs from other construction types; while there is substantial redundancy

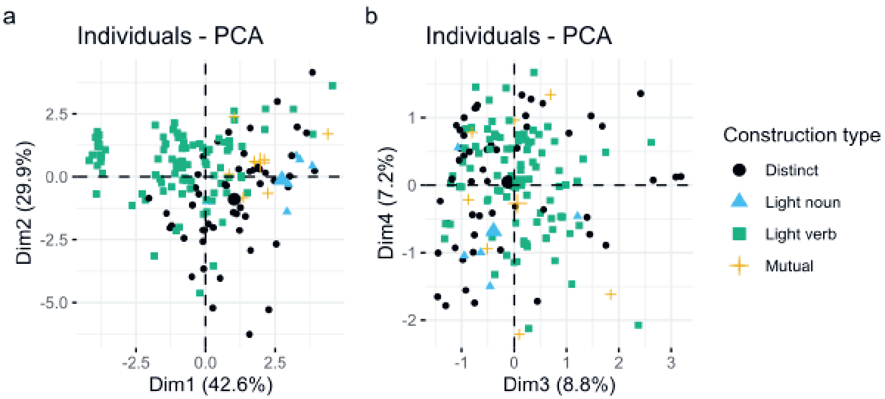

Fig. 9: Map of the different N-V combinations on the first four principal dimensions.

between these measures, they do have contributions independent of bidirectional association, and noun entropy may have a contribution independent of all other factors. I conclude that to characterise the statistical distributional properties of LVCs, it is important to look beyond classic association measures, and examine measures that zero in on the differences between the nominal and verbal slots. It is hoped that future LVC detection methods will add unidirectional association and slot flexibility measures.

Although the points raised in this paper are limited to Tibetan, they may also apply to light verb detection more generally. For example, English also has N-V collocations where the verb predicts the noun (*part ways*), the noun and the verb have similar semantic contributions (*sing a song, fire a shot*) or where the two elements' contributions are relatively distinct (*slice carrots*). It is also hoped that even more potentially useful characteristics of LVCs, such as semantic domains (Taslimipoor et al., 2012; Vaidya et al., 2016; Singh et al., 2016), distribution of interveners, argument structure (Simon, 2011; Singh et al., 2016), and register (Tibetan LVCs are considered more informal than Mutual and Distinct constructions with similar meanings; Geissler, 2018, 11) may be considered for the Tibetan LVC detection toolkit in the future. Finally, as this paper focuses on high-frequency combinations, it rests on the general assumption that LVCs are conventionalised combinations; however, it is possible that there may be LVCs that are created on an ad hoc basis. I leave it to future work to investigate this issue in more detail, which may require larger datasets and more precise measurements of productivity than have been explored in this paper.

Acknowledgments

This chapter was written as part of my PhD thesis at UC Santa Barbara. I would like to thank my committee, Stefan Th Gries, John W DuBois, Wendy Meiring and Satoko Shirai as well as members of UCSB's CEILing group for their valuable comments.

Bibliography

Agha, Asif. 1993. *Structural form and utterance context in Lhasa Tibetan: grammar and indexicality in a non-configurational language*. New York: Peter Lang.

Aikhenvald, Alexandra and Robert M. W. Dixon. 2006. *Serial verb constructions: a cross-linguistic typology*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199279159.001.0001.

Bailey, Geoff and Christopher E. Walker. 2004. *Lhasa verbs: a practical introduction*. Lhasa: Tibetan Academy of Social Science.

Bartee, Ellen Lynn. 2007. *A grammar of Dongwang Tibetan*. University of California, Santa Barbara.

DeLancey, Scott. 1991. The origins of verb serialization in modern Tibetan. *Studies in Language* 15(1). 1–23. https://doi.org/10.1075/sl.15.1.02del.

Denwood, Philip. 1999. *Tibetan*. Amsterdam: John Benjamins.

Desagulier, Guillaume. 2016. A lesson from associative learning: asymmetry and productivity in multiple-slot constructions. *Corpus Linguistics and Linguistic Theory* 12(2). 173–219. https://doi.org/10.1515/CLLT-2015-0012.

Dras, Mark and Michael Johnson. 1996. Death and lightness: using a demographic model to find support verbs. In *International Conference on the Cognitive Science of Natural Language Processing (5th: 1996)*, Dublin: Dublin City University Natural Language Group.

Dunning, Ted. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1). 61–74.

Esukhia. 2023. *botok*. https://github.com/OpenPecha, last access 27/9/2024.

Evert, Stephanie. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Doctoral dissertation: Universität Stuttgart.

Geissler, Christopher. 2018. Phonological koinéization in Kathmandu Tibetan. In Gillian Gallagher, Maria Gouskova and Sora Yin (eds.), *Proceedings of the Annual Meetings on Phonology*, 1–12. Washinton D.C.: Linguistic Society of America. https://doi.org/10.3765/amp.v5i0.4242.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next .... *International Journal of Corpus Linguistics* 18(1). 137–166. https://doi.org/10.1075/ijcl.18.1.09gri.

Gries, Stefan Th. 2022. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis: Journal in English Lexicology* 19(3). 1–23. https://doi.org/10.4000/lexis.6231.

Hackett, Paul. 2003. *A Tibetan verb lexicon: verbs, classes, and syntactic frames*. Ithaca: Snow Lion Publications.

Hill, Nathan W. 2010. *A lexicon of Tibetan verb stems as reported by the grammatical tradition*. Munich: Bayerische Akademie der Wissenschaften.

Hill, Nathan W. and Marieke Meelen. 2017. Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics* 16(2). 64–86. https://doi.org/10.5070/H916234501.

Hoshi, Izumi. 2003. *Gendai Chibettogo Doushi Jiten [A verb dictionary of the modern spoken Tibetan of Lhasa]*. Tokyo: Tokyo University of Foreign Studies.

Kjellmer, Goran. 2014. A mint of phrases. In Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics*, 123–139. London: Routledge.

Lai, Ryan Ka Yau and Michelle Man-Long Pang. 2023. Rethinking the description and typology of Cantonese causative–resultative constructions: a dynamic constructionist lens. *Languages* 8(2). 1–48. https://doi.org/10.3390/languages8020151.

Lessan Pezechki, Homa and Nicolas Tournadre. 2014. La question du sujet et les verbes support: Les cas du persan et du tibétain. Paper presented at Du sujet et de son absence. Le Mans, France.

Michelbacher, Lukas, Stefan Evert and Hinrich Schütze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* 7(2). 245–276. https://doi.org/10.1515/cllt.2011.012.

Michelbacher, Lukas, Stephanie Evert and Hinrich Schütze. 2007. Asymmetric association measures. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing (ranlp)*, 367–372. Amsterdam: John Benjamins.

Monlam, Lobsang. 2016. *Monlam Tibetan-English Dictionary*. https://github.com/iamironrabbit/monlam-dictionary, last access 6/3/2023.

Nagy, István and Veronika Vincze. 2011. Identifying verbal collocations in Wikipedia articles. In Ivan Habernal and Václav Matoušek (eds.), *Text, Speech and Dialogue*, 179–186. Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23538-2_23.

Randall, Michael Gordon. 2016. *The properties of Lhasa Tibetan verbalizers*. Master's thesis: Payap University, Chiang Mai.

Schmidt, Dirk. 2019. A speech corpus of Dharamsala Tibetan. 10.17613/r9x9f-a3174. Paper presented at HLS25: 25th Himalayan Languages Symposium. Sydney, Australia.

Simon, Camille. 2011. *Derivation causative en tibétain (Lhasa)*. Master's thesis: Université de Provence, Aix-en-Provence.

Singh, Dhirendra, Sudha Bhingardive and Pushpak Bhattacharyyaa. 2016. Detection of compound nouns and light verb constructions using IndoWordNet. In Verginica Barbu Mititelu, Corina Forascu, Christiane Fellbaum and Piek Vossen (eds.), *Proceedings of the 8th global WordNet conference (GWC)*, 404–410. Weesp.

Stefanowitsch, Anatol and Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43. https://doi.org/10.1515/cllt.2005.1.1.1.

Steinert, Christian. 2023. *Tibetan-English dictionary application*. https://github.com/christiansteinert/tibetan-dictionary, last access 6/3/2023.

Stevenson, Suzanne, Afsaneh Fazly and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In Takaaki Tanaka, Aline Villavicencio, Francis Bond and Anna Korhonen (eds.), *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 1–8. Barcelona: Association for Computational Linguistics.

Tan, Yee Fan, Min-Yen Kan and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In Paul Rayson, Serge Sharoff and Svenja Adolphs (eds.), *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*, 49–56. Trento: Association for Computational Linguistics.

Taslimipoor, Shiva, Afsaneh Fazly and Ali Hamzeh. 2012. Using noun similarity to adapt an acceptability measure for Persian light verb constructions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the eighth international conference on language resources and evaluation (lrec'12)*, 670–673. Istanbul: European Language Resources Association (ELRA).

Tournadre, Nicolas. 1996. *L'ergativité en tibétain: approche morphosyntaxique de la langue parlée*. Louvain: Peeters.

Tournadre, Nicolas and Sangda Dorje. 1998. *Manuel de tibétain standard: langue et civilisation: introduction au tibétain standard (parlé et écrit) suivie d'un appendice consacré au tibétain littéraire classique*. Paris: L'Asiathèque-Maison des langues du monde.

Tu, Yuancheng and Dan Roth. 2011. Learning English light verb constructions: contextual or statistical. In Valia Kordoni, Carlos Ramisch and Aline Villavicencio (eds.), *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, 31–39. Portland: Association for Computational Linguistics.

Vaidya, Ashwini, Sumeet Agarwal and Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. In Yuji Matsumoto and Rashmi Prasad (eds.), *Proceedings of Coling 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1320–1329. Osaka: Coling.

Vincze, Veronika, István Nagy and Gábor Berend. 2011. Detecting noun compounds and light verb constructions: a contrastive study. In Valia Kordoni, Carlos Ramisch and Aline Villavicencio (eds.), *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 116–121. Portland: Association for Computational Linguistics.

Zemp, Marius. 2018. *A grammar of Purik Tibetan*. Leiden: Brill.

Zhào, Wéinà, Lin Li, Huìdān Liú and Jiàn Wú. 2016. Tibetan trisyllabic light verb construction recognition. *Himalayan Linguistics* 15(1). 137–148. https://doi.org/10.5070/H915130102.

Zhào, Wéinà, Lín Lǐ, Huìdān Liú, Pǔbùdùnzhū and Jiàn Wú. 2015. Automatic extraction of trisyllabic verb phrases in Tibetan. *Journal of Chinese Information Processing* 29(3). 196–200.