Why we need asymmetric measures to classify multi-word expressions: The case of Tibetan light verb constructions

Ryan Ka Yau Lai
University of California, Santa Barbara
kayaulai@ucsb.edu

1 Introduction

In recent years, linguistics has devoted much attention to the description and processing of *multi-word expressions*, conventionalised strings of multiple, possibly non-contiguous words, that function semantically like a single lexical item. An important subtype of MWEs is light verb constructions (LVCs), which have a semantically light verb and a semantically heavy lexical item (usually a nominal) that carries most of the predicational information, e.g. *give (it) a try* or *take a nap* in English. This study examines the statistical properties of LVCs in modern Tibetan, where they exhibit special grammatical properties and are highly ubiquitous (Randall 2016), including expressions as basic as བེད་སྤྱོད་གཏོང་ *bed.spyod gtong* (use + send = 'to use') and གོམ་པ་རྒྱག་ *gom.pa rgyag* (step + strike = 'to walk'). Tibetan LVCs are of particular interest as they are relatively overlooked by linguistic resources: most dictionaries other than Bailey & Walker (2004) offer little grammatical and usage information. A better understanding of the statistical properties of LVCs will facilitate the extraction of LVCs from large-scale corpora, and hence the construction of further lexical resources.

Much research in computational corpus linguistics has examined automatic methods to extract LVCs from large-scale corpora. Particularly versatile are those based on statistical measures of co-occurrence, which are easily generalisable across languages and do not require extensive existing lexical resources. Most commonly, researchers use bidirectional association measures like pointwise mutual information (PMI) (e.g. Tan et al. 2006), which measure how much the verb and nominal prefer to co-occur with each other. For Tibetan, Zhào et al. (2015, 2016) have also used the entropy of surrounding tokens to measure the diversity of contexts where the combination appears. Such systems usually assume that the higher the value of these measures, the greater the chance of an N-V combination being an LVC.

A potential disadvantage of these measures is that the noun and verb are treated equally: They do not separate the noun's attraction to the verb from the opposite attraction. Thus, they may be unable to distinguish LVCs from other types of noun-verb combinations where the verb is *not* semantically light and/or the noun is *not* semantically heavy. Tibetan has at least three such types of noun-verb collocations that do not fit in the LVC description. Firstly, the noun may be semantically lighter than the verb. For example, in ཤོད་ཁོག་སྡུག་ *grod.khog ltogs* (stomach + be hungry = 'to be hungry'), 'hungry' already contains most of the meaning; 'stomach' does not add much, as that is the only body part that experiences hunger. These will be called light noun constructions (LNC), although the noun is typically not as light as light verbs. Secondly, the noun and verb may carry similar information, such as རྐུན་མ་རྐུ་ *rkun.ma rku* (thief + steal = 'to steal'), where the idea of stealing is conveyed by both parts of the construction, and hence both parts are easily predictable from the other. These will be called MUTUAL constructions. Finally, both elements may carry distinct, substantial meanings, e.g. དེབ་ཀློག་ *deb.klogs* (book + read 'to read a book'); these will be called DISTINCT constructions.

This study examines the statistical properties of LVCs in comparison to the other three types of N-V combinations. In additional to bidirectional association and context entropy, I examine several measures of unidirectional association, as well as the productivity of one slot given the other. These measures are *asymmetric*, treating the noun and verb differently, and thus are potentially better at differentiating LVCs from non-LVCs. For each measure, this study examines how LVCs differ from the other three types, and I propose suggestions to improve MWE extraction based on these findings.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 302-306.
Irvine, California, June 27-29, 2024

2 Methodology

2.1 Extraction of noun-verb pairs.

I first extracted from the Nanhai corpus (Schmidt 2020) pairs of verbs with non-case-marked nouns (excluding potential LVCs with case-marked nouns as they are uncommon). This was a four-step process, and at each step, common false positives and false negatives were identified and filtered out or brought back in. Firstly, verbs were identified using Hill (2010) and stemmed. Secondly, nouns were identified with the Monlam dictionary (Lobsang Monlam et al. 2016) and POS tags from a Classical Tibetan corpus (Hill & Meelen 2020). Thirdly, the text was separated into segments demarcated by the Tibetan punctuation mark *shad*, and for each verb, preceding content in the same segment was scanned. If a noun appeared before the verb, with no verb, noun, case marker, quotative or unknown word between the two, the pair was extracted as a candidate noun-verb combination. Finally, verbs were lemmatised. Remaining false positives were flagged and removed, and the 183 extracted combinations with token frequency >10 were annotated for the four construction types.

2.2 Measures computed

Based on the final list of combinations, various measures of co-occurrence were calculated as in Table 1 using R (R Core Team 2013). The bidirectional association and external flexibility are symmetric measures, while the unidirectional association and productivity measures are asymmetric: they do distinguish between the different roles of the noun and the verb.

3 Results

For bidirectional and unidirectional association as well as external flexibility measures, Wilcoxon's rank-sum tests were used to compare LVCs' values against the other three constructions, with Holm-Bonferroni correction at $\alpha = .05$ within each set of three comparisons. Since productivity measures belong to particular nouns or verbs instead of pairs, differences between the constructions were tested using a Dirichlet regression model with LVC as reference category, each productivity measure as predictor, and the proportions of each construction type as the outcome variable; nonzero coefficients indicate that the construction types differ with respect to the productivity measure.

Bidirectional association	● PMI between N and V ● G^2 value (Dunning 1994)
External flexibility	● Entropy of previous and next words (Zhào et al. 2016)
Uni-directional association	● Conditional surprisal of V given N and *vice versa* ● Rank of G^2 for Vs for each N, and *vice versa* (Michelbacher et al. 2011) ● Normalised KLD of the V's distribution given the N compared to the V's overall distribution, and *vice versa* (Gries 2022) ● ΔP , i.e. probability of getting the N given verb minus probability of getting the N given any other verb, and *vice versa* (Gries 2013)
Productivity	● Number of V types that appear with each N and *vice versa* ● Entropy of the V given then N and *vice versa* (Gries 2022)

Table 1: The measures examined in this study.

3.1 Symmetric measures

For bidirectional association (Figure 1), no significant difference between G^2 values was found between LVCs and any of the non-LVC constructions (p = 0.0467 for LN, 0.0189 for DISTINCT, 0.4013 for MUTUAL). PMI values for LVCs were significantly *lower* for LVCs compared to MUTUAL and LNCs, contrary to the traditional assumption that higher PMIs are indicative of LVC status (p = 0.00280 for LN, 0.37182 for DISTINCT, 0.00030 for MUTUAL).

While there was a significant difference in preceding and following token entropy between LVCs and DISTINCT constructions (preceding tokens: p = 0.1653 for LN, 0.0045 for DISTINCT, 0.9247 for MUTUAL; following tokens: p = 0.55 for

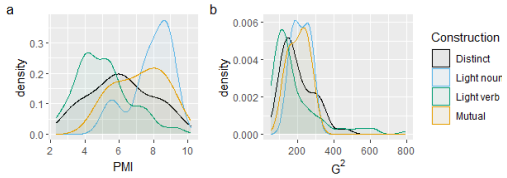

Figure 1: Kernel density estimates (KDEs) of the distributions of (a) PMIs and (b) G^2s of the four constructions.

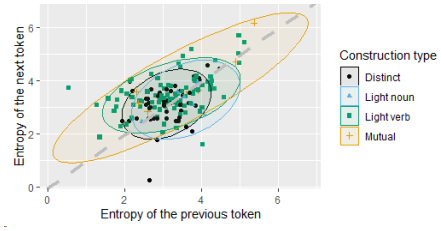

Figure 2:.Scatterplot of context entropies of different construction. The grey diagonal indicates points where previous and next word entropy are equal.

LN, 0.000014 for DISTINCT, 0.90 for MUTUAL), it is visually very small (Figure 2).

3.2 Asymmetric measures

On the contrary, unidirectional association (Figure 3) show much more promise. As ΔP and normalised KLD are close to monotonic functions of conditional surprisal for this dataset, statistical analysis focused on surprisal and G^2 rank. Nouns are clearly much more attracted to verbs in LVCs than in the other three construction types (surprisal: $p = 0.00040$ for LNC, 6.0×10^{-7} for DISTINCT, 6.3×10^{-5} for MUTUAL; G^2 rank: 0.00024 for LNC, 1.3×10^{-11} for DISTINCT, 0.00023 for MUTUAL). The verbs' attraction to the noun are also significantly different between LVCs and DISTINCT constructions (surprisal: $p = .081$ for LNC, 1.4×10^{-7} DISTINCT, .44 for MUTUAL; G^2 rank: 0.2623 for LNC, 0.0016 for DISTINCT, 0.5598 for

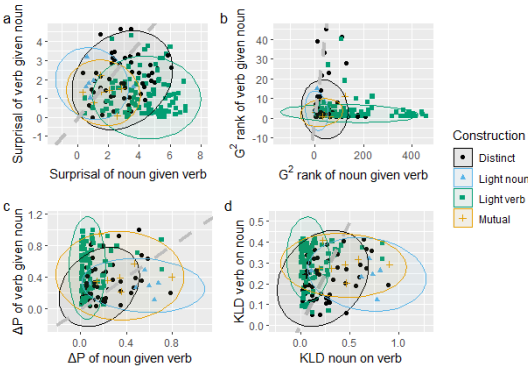

Figure 3: Scatterplot of unidirectional association measures for the four constructions. The *x*-axis gives the nouns' attraction to the verbs, and the *y*-axis gives the reverse attraction.

MUTUAL); visual inspection suggests strong differences for LNCs too.

As for productivity measures, the productivity of the verb slot given the noun was clearly greater for LNCs and DISTINCT than for LVCs (type frequency: $p = 2.19 \times 10^{-7}$ for LNC, .0147 for DISTINCT, .956 for MUTUAL; entropy: 1.82×10^{-12} for LNC, .00427 for DISTINCT, .314 for MUTUAL); the lack of a comparable result for MUTUAL constructions is unsurprising given the semantic specificity of these constructions. Visual inspection also clearly suggests that the productivity of the noun slot given the verb may be higher for LVCs than the other three (though none of the comparisons reached significance, likely because of the small number of available verbs resulting in a low sample size).

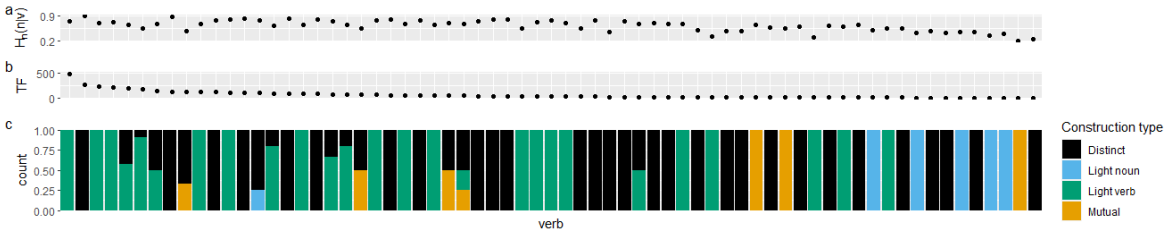

Figure 4: (a) Normalised entropies of the noun slot for each verb, (b) number of noun types appearing with each verb, and (c) proportion of the four constructions by verb.

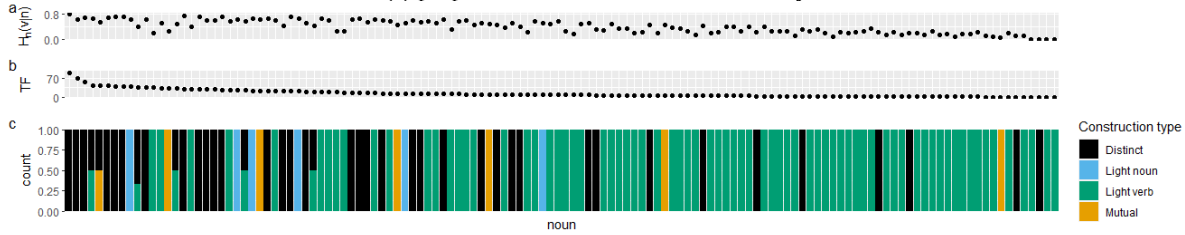

Figure 5: (a) Normalised entropies of the verb slot for each noun, (b) number of verb types appearing with each noun, and (c) proportion of the four constructions by noun.

All these results are expected from the semantic properties of the various construction types. Light verb nominals tend to strongly prefer a few verbs and disprefer others; light verbs, being semantically light, are compatible with a wide variety of nominals without specifically preferring a subset of them.

3.3 Combining measures

To determine the level of redundancy among the measures, a principal components analysis was conducted on all the measures (except the context entropies, and replacing G^2 with χ^2 to avoid undefined values). It is found that three dimensions suffice to represent 87.5% of variation in the data. The locations of the various constructions within the first four principal dimensions are shown in Figure 6.

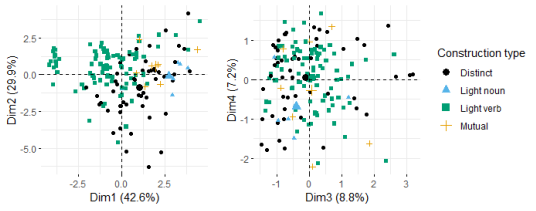

Figure 6: Scatterplot of the constructions in the first four principal dimensions of the PCA.

Dimension 1 (42.6% of the variance) correlates primarily with high bidirectional association, and low values for measures associated with lighter verbs and heavier nouns; LVCs typically have low values and LNCs high values, with the other two constructions in between. Dimension 2 (29.9%) correlates with high bidirectional association and low values for measures associated with lighter nouns and heavier verbs, so LVCs have values somewhat higher than the rest, with DISTINCT constructions being particularly low. 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 correlations between measures and principal components reveals with three 'clusters' of measures associated with light nouns/heavy verbs, heavy verbs/light nouns and bidirectional association respectively; however, the noun's entropy given the verb appears to act independently of all these.

	Precision	Recall
PC1	1	.667
PC2	.5625	1
PC3	.615	.889
Any 2-3 PCs	.818	1

Table 2: Precision and recall of the GAM using different principle dimensions.

A classifier based on a logistic generalised additive model in `mgcv` (Wood 2014) with no interactions was used to see how well the combination of these measures serve to predict whether a frequent noun-verb combination is a light verb construction or not. 10% of the data was used as the holdout set. As shown in Table 2, considerable classification accuracy is achieved using any two (or all 3) of the principal components.

4 Discussion and conclusion

This study found that traditional symmetric measures used for extracting light verb constructions do not necessarily work well for Tibetan. Rather, asymmetric measures, including productivity measures of specific slots and unidirectional attraction measures, are more effective in distinguishing light verb constructions from other noun-verb collocations.

Although the case study was done on Tibetan LVCs, the same principles likely also apply to any studies aiming to extract specific categories of MWEs with known semantic (a)symmetries in any language. For example, English too has V-N sequences where the verb is predictive of the noun (*part ways*), where the two elements are semantically similar (*sing a song*, *fire a shot*) or where they have relatively distinct contributions (*slice carrots*); asymmetric measures may also help to differentiate LVCs from these. Moreover, other MWE types from LVCs may also benefit from asymmetric measures: for example, adposition-relational noun combinations (e.g. *on top*) may be expected to be more symmetric than other adposition-noun combinations (e.g. *on time*), and thus benefit from using co-occurrence statistics that can capture this symmetry, rather than using only single association measures like the PMI. Thus, I hope these findings can also be applied to other contexts in the future.

References

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

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

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.

Gries, Stefan Th. 2022. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis: Journal in English Lexicology* 3 (19).

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

Hill, Nathan W. & Marieke Meelen. 2017. Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics* 16(2). 64–86.

Lobsang Monlam et al. (2016). *Monlam Tibetan-English Dictionary.* https://github.com/iamironrabbit/monlam-dictionary. Accessed 6th March 2023.

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

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

Schmidt, Dirk. 2020. A speech corpus of Dharamsala Tibetan. Paper presented at HLS25: 25th Himalayan Languages Symposium. Sydney, Australia. https://doi.org/10.17613/BNZT-XD51.

Tan, Yee Fan, Min-Yen Kan & Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*.

Wood, Simon & Fabian Scheipl. 2014. `gamm4`: Generalized additive mixed models using mgcv and lme4. *R package version 0.2-3*.

Zhào, Wéinà [赵维纳], Lín Lǐ [李琳], Huìdān Liú [刘汇丹], Pǔbùdùnzhū [普布顿珠] & Jiàn Wú [吴健]. 2015. Automatic extraction of trisyllabic verb phrases in Tibetan [藏语三音动词短语自动抽取研究]. *Journal of Chinese Information Processing* [中文信息学报] 29(3). 196–200.

Zhao, Weina, Lin Li, Huidan Liu & Jian Wu. 2016. Tibetan trisyllabic light verb construction recognition. *Himalayan Linguistics* 15(1).