More frequent verbs are associated with more diverse valency frames: Efficient language design at the lexicon-grammar interface

Anonymous ACL submission

Abstract

A substantial body of work has provided evidence that the lexicons of natural languages are organized to support efficient communication. However, existing work has largely focused on word-internal properties, such as Zipf's classical observation that more frequent words are optimized in form to minimize communicative cost. Here, we investigate the hypothesis that efficient lexicon organization also affects valency, or the combinations and orders of additional words and phrases a verb selects for in a sentence. We consider two measures of valency diversity for verbs: *valency frame count* (VFC), the number of distinct frames associated with a verb, and *valency frame entropy* (VFE), the average information content of frame selection associated with a verb. Using data from 79 languages, we provide evidence that more frequent verbs are associated with a greater diversity of valency frames, suggesting that the organization of valency helps optimize communicative efficiency. We discuss our findings in relation to classical findings such as the principle of least effort, as well as implications for theories of valency and efficient language design.[1]

1 Introduction

The idea that functional pressures shape the organization of natural language lexicons has a long tradition, going back at least to Zipf's law of abbreviation: the length of the word is inversely related to the frequency of its use, motivated by the "principle of least effort" (Zipf, 1935, 1949). Further evidence has been provided by a functionalist line of research studying the emergence of the lexicon from the frequency of use (e.g. Bybee, 1998; Hopper and Bybee, 2001) as well as more recent work that formalizes such ideas using information theory (e.g. Regier et al., 2015; Wu et al., 2019; Mahowald et al., 2022; Pimentel et al., 2021). Common across this literature is the idea that the lexicon reflects a pressure for efficient language use under the constraints posed by human cognition. Evidence has been adduced from domains such as word lengths (Piantadosi et al., 2011; Pimentel et al., 2023), morphological complexity (Wu et al., 2019), orthographic forms (Mahowald et al., 2018), and the partitioning of semantic space into word meanings (e.g. Regier et al., 2015; Kemp et al., 2018; Zaslavsky et al., 2018).

However, a key aspect of the lexicon has gone unaddressed in this literature: words are not used in isolation but systematically integrated within the broader structure of a sentence. Verbs in particular play a key role in determining the syntactic and semantic structure of a sentence, and the study of *valency* features prominently in theoretical linguistics (e.g. Chomsky et al., 1970; Levin and Rappaport Hovav, 2005). In broad strokes, the *valency* of a verb indicates the type and number of dependents it takes. Certain features of valency are shared across many but not all languages, e.g., a majority of verbs in many languages minimally requires a subject (Chomsky, 1982; McCloskey, 1994); and a transitive category of verbs may take on one or more objects in addition (Bowers, 2002). Typological literature has documented the many ways in which valency differs across verbs within a language, and across languages (e.g. Hartmann et al., 2013).

Fig. 1 shows how the sentences equivalent to English *The teacher helped the children with the homework* are rendered differently in several languages, with dependency relations of verb annotated according to Universal Dependencies (UD, v.2 Nivre et al., 2020) guidelines. We observe both cross-lingual similarities and differences: for example, there is a subject (UD relation `nsubj`), an object (`obj`) and an oblique dependent (`obl`) in both English and Finnish, but all three dependents are

[1] Code for reproducing our results is available under GNU GPL Version 2 at https://anonymous.4open.science/r/verbal-valency-ud-8B7E

marked for case in Finnish (nominative, partitive and inessive, respectively). Other languages use other combinations of relation types to express the same meaning; for instance, the verb has only two dependents in Japanese and Chinese: in Japanese the verb takes *homework* as `obj` and in Chinese the sentence must be formulated with another verb as a clausal complement (`ccomp`).

Within a language, different verbs allow different *valency frames*: specific configurations of morphosyntactic features they appear with. For example, Fig. 2 shows some other valency frames observed with the English verb *help* (also known as its diathesis alternations). There is general consensus that the syntactic frames can at least in part be predicted by semantics (e.g. Fillmore, 1968; Levin, 1993), but divergent explanations of the findings have been offered (Gropen et al., 1989; Goldberg, 1992). A substantial body of work has also classified verbs within a language based on their diathesis alternations (e.g. Levin, 1993).

Here, we investigate whether the valency properties of verbs show signatures of efficient organization. We argue that considerations of communicative efficiency predict that verbs will systematically differ in the diversity of valency frames they are associated with, so that high-frequency verbs will be associated with a greater diversity of frames. A range of theoretical considerations converge on this prediction. For example, there is existing evidence of a correlation between semantic ambiguity and verb frequency – more frequent verbs tend to be more ambiguous semantically (Hoffman et al., 2013), and might thus have more frames associated with their different senses. From a comprehension perspective, this kind of pattern ensures frames are predictable when verbs are not, reducing spikes in information density. We argue that similar considerations apply from the perspective of language production and learning.

We operationalize valency at a lexeme-level, and propose two novel information-theoretic metrics of the diversity of valency frames: (i) **valency frame count** (VFC) that measures the number of distinct frames associated with a verb; and (ii) **valency frame entropy** (VFE) that measures the average surprisal of valency frames as conditioned by the verb, determined by the diversity of frames associated with that verb. This allows us to posit and test hypotheses regarding the correlation between these metrics and the diversity of valency frames in experiments and show that effects predicted by learnability and efficiency considerations shape valency systems of languages cross-linguistically.

2 Background and Motivation

The Notion of Valency The notion of valency, and closely related notions such as *argument structure*, *subcategorization* and *diathesis alternation*, have received a variety of formalizations across formal linguistic theories (e.g. Tesnière, 1959; Chomsky, 1965; Fillmore, 1982; Levin, 1993; Goldberg, 1995). While many theories make a formal distinction between arguments and adjuncts (Chomsky et al., 1970; Ackema, 2015), others treat them in a similar fashion (van Noord and Bouma, 1994). For this study, we abstract away from such theoretical questions, and adopt a broad definition of verb valency within UD, a dependency grammar framework. In doing so, we aim for a broad descriptive basis for cross-linguistic comparison, as UD classifies dependency relations into cross-linguistically meaningful types (cf. Figures 1–2); pragmatically, this also allows applicable operationalization on large-scale cross-linguistic usage data available as UD treebanks. We formalize as *valency* the set of such relation types of dependents a verb co-occurs with, and a **valency frame** as a set of UD relations corresponding to dependents of a verb (see § 4.1).

Computational Studies of Valency A key line of computational research on valency has created datasets of the valency frames associated with verbs, such as FrameNet (Baker et al., 1998; Fillmore and Baker, 2015) and VerbNet (Kipper-Schuler, 2005; Kipper et al., 2006, 2008). Mu et al. (2017) automatically classifies verbs in terms of their valency as listed in such resources to recover classifications from the theoretical literature (Levin, 1993). Cross-linguistically, Baker and Lorenzi (2020); Ellsworth et al. (2021) explores the alignment of frames based on FrameNet. In contrast to their approach, our operationalization of valency frame is morphosyntactic in nature, directly in terms of UD relations. This allows us to match valency frames with large scale usage data across many languages and for the full set of verbs appearing in any given treebank – something that is not available when operationalizing frames using curated lexical resources such as FrameNet.

Theories of Communicative and Processing Efficiency One justification for efficient organization

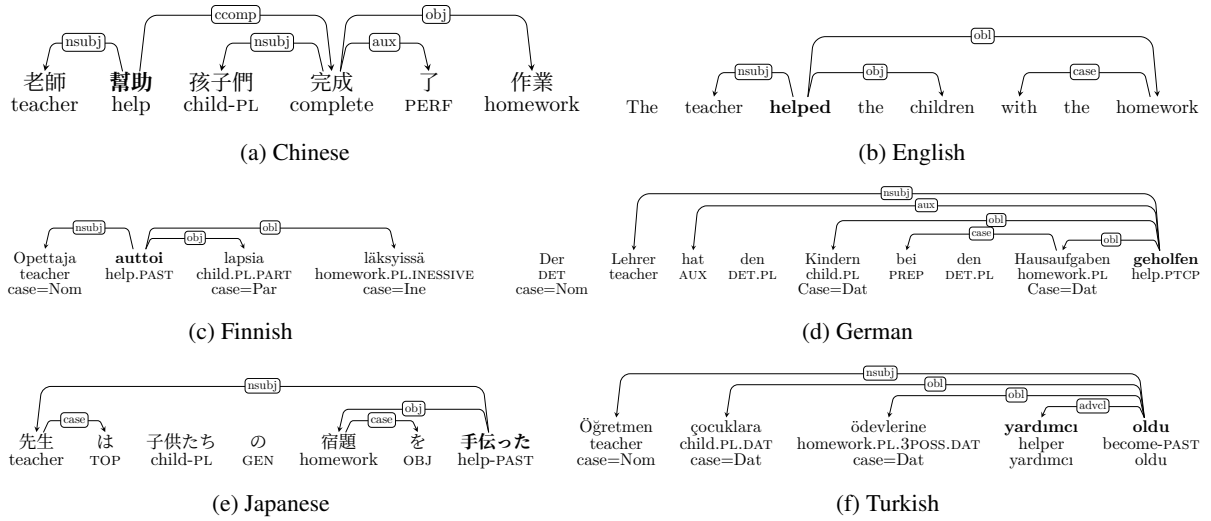


Figure 1: Example sentences in 6 languages showing relevant dependency relations. In each language, the verb expressing "helping" is associated with a different set of UD relation types linking it to its dependents.

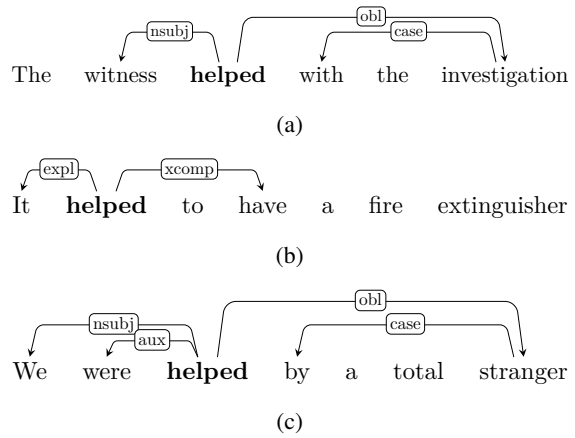


Figure 2: Further examples of the verb "help" in English, showing its diathesis alternations.

of the lexicon comes from *learnability* considerations. For instance, it has been suggested that infrequent irregular verbs are poorly acquired, leading to morphological irregularity being focused on high-frequency words (e.g. Bybee, 1998; Wu et al., 2019). Similar considerations apply to valency: When learning from a finite sample, a learner should be able to acquire a larger variety of frames for frequent verbs, as more data is available for them. Thus, if learnability affects the organization of valency, one expects that infrequent verbs should be associated with smaller diversity of frames.

Another key justification for efficient organization of the lexicon (and language in general) comes from online processing, in both language production and comprehension. On the *production* side, accessibility, i.e. the ease of retrieval, has been shown to play a key role in shaping language form and structure (Kathryn Bock and Warren, 1985; MacDonald, 2013). To the extent that verbs may be planned in advance of their arguments in sentence production (e.g. Momma and Ferreira, 2019), reducing the diversity of frames associated with a verb should increase the accessibility of any of these frames. Such a pressure should be stronger for the long tail of low-frequency verbs, as producers will be less attuned to their production. Conversely, being compatible with a larger number of frames could make a verb more versatile and therefore encourage its use.

On the *comprehension* side, we expect consequences for sentence interpretation as well as the distribution of information density. Psycholinguistic evidence shows that argument structure information is used in online comprehension to eliminate implausible interpretations (Boland et al., 1995). The more frequent verbs, being more ambiguous or vague (Hoffman et al., 2013), would require additional information to disambiguate or narrow down the meaning. Additionally, the Uniform Information Density (UID) hypothesis (Fenk and Fenk-Oczlon, 1980; Levy and Jaeger, 2006) posits that language speakers prefer a more even distribution of surprisal values across utterances in order to maximize but not overload the capacity of the communication channel. Less frequent verbs will, on average, have high surprisal; high entropy in the valency frames associated with it would lead to an undesirable spike in overall surprisal.

3 Data and Methodology

3.1 Data

We draw on Universal Dependencies (UD) as the source of data. It provids as a cross-linguistically consistent system for annotating morphosyntactic information within a dependency grammar framework (de Marneffe et al., 2021). We use the UD v2.11 release, and include all languages with at least 10,000 tokens across treebanks. We exclude L2 learner corpora and code-switched corpora. Korean corpora are excluded due to the lack of verb lemmatization, which is necessary for empirically estimating valency. In total, 79 languages out of the 138 languages remained.

3.2 Quantifying Diversity of Frames and Verbs

We hypothesize a positive relation between a verb's frequency and the diversity of frames that it occurs with. In order to test this, we propose two measures of the diversity of frames associated with a verb.

The simplest metric of the diversity of frames associated with a verb is **valency frame count** (VFC): the number of distinct frames that a verb occurs with in the corpus.

By definition, VFC gives equal weight to more or less frequent frames. As discussed in §2, information-theoretic models of language comprehension such as UID predict an effect of the *entropy* of the frames, conditioned on the verb. Our second measure formalizes this. Let R be the set of UD relations under consideration. Let F be a random variable ranging over subsets of R – that is, over frames. Let V be a random variable ranging over the verb lemmata v in a language. A dependency treebank defines, for each verb v, a probability distribution $P(f|V = v)$ over different frames proportional to their frequencies of occurrence. We formalize the **valency frame entropy** (VFE) as $H[F|V = v]$, the entropy of the frame F conditioned on a verb v:

$$-\sum_{f \subseteq R} P(f|V = v) \log_2 P(f|V = v) \quad (1)$$

A related measure, but focusing on only complements, was used as a predictor of English online processing cost in Linzen et al. (2013); Linzen and Jaeger (2015). Valency frame entropy equals $\log(\text{VFC})$ in the special case where all frames that appear with a verb do so at equal frequencies. However, keeping VFC constant, VFE is lowered when the frames appear at more skewed frequencies.

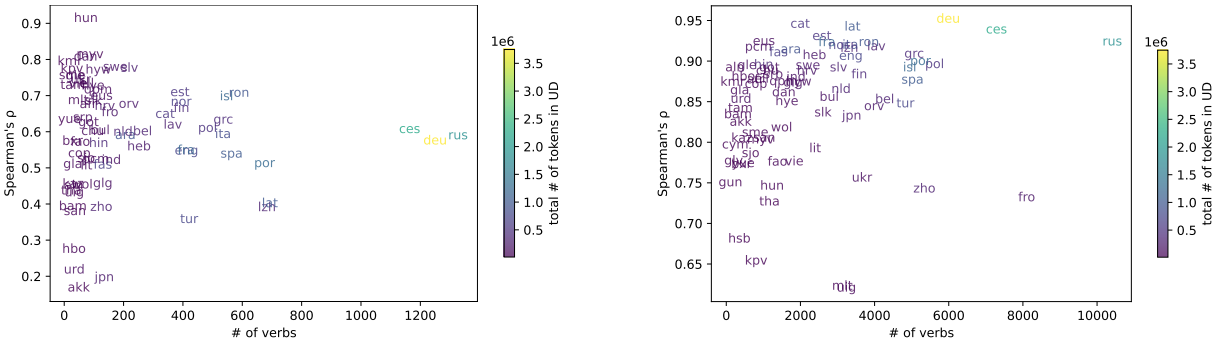
4 Experiment

4.1 Dependency Relations

As we take a broad view on the notion of valency, we correspondingly include a broad range of UD relations (Figure 6). Definitions of valency vary across the literature, but conveniently UD classifies the relevant set of relations. We included the UD v2 relations defined by UD as *core arguments* and *non-core dependents*. Core arguments, including – among others – subjects and objects are most clearly associated with argument structure in typical theories of valency and syntax. Non-core dependents are other possible dependents of a clausal predicate, including – among others – oblique nominals and adverbial modifiers. UD further includes *nominal dependents* – dependents of nouns – which are by definition of no relevance to our study. Finally, UD also has the categories *coordination*, *headless*, *loose*, *special*, and *other*; we do not consider any of these to be relevant: for instance, coordination (*conj* and *cc*) defines a symmetric relationship rather than a hierarchical dependency structure; and *special* lacks cross-lingual applicability. We examine correlations estimated from various subsets in Figure 5.

4.2 Controlling for Confounds via Cross-Entropy and Subsampling

Estimating count and entropy measures on finite samples is a difficult problem, and popular estimators for such quantities are generally biased (e.g. Paninski, 2003). As our goal here is to test a hypothesized relationship, we aim for estimators that avoid any bias towards false positives in the directions of our hypothesis – that is, estimators that lead to conservative estimates for the testing of our hypothesis. Indeed, testing our hypothesized relationship using naive estimators of VFC and VFE faces a **circularity confound**: if there are few observations for a verb in a corpus, the number of distinct frames observed with it will necessarily be small, leading to an underestimation bias for entropy and frame count. As precisely these verbs are expected to have low frame entropy and low frame count under our hypotheses, it may lead to spurious correlations. Stated differently, low-frequency verbs might have a low measured diversity of frames simply due to lack of observations. We use two



(a) results using subsampling

(b) results not using subsampling

Figure 3: Scatter plots showing languages by their ISO codes with the number of verbs across languages on the x-axis, and Spearman's rank correlation between VFE and frequency rank of verbs on the y-axis. Color shows the total number of tokens in UD. Results use subsampling in subfigure 3a and full data in subfigure 3b; all use cross-entropies. Note that the y-axes have different ranges in (a) and (b) to maximize visibility of the individual languages.

strategies to eliminate this confound.

Estimation via Cross-Entropy Our first strategy is to empirically estimate the valency frame entropy using the cross-entropy. That is, for each verb, we randomly split its observations in the corpora into two halves, resulting in two empirical distributions $P(F|V = v)$ and $P(F'|V = v)$. The cross-entropy between them, conditioned on a verb v, is:

$$-\sum_{f \subseteq R} P(xf|V = v) \log_2 P'(f|V = v) \qquad (2)$$

We use Laplace smoothing with $\alpha = 1$ over the full set of frames to ensure well-definedness.

How Cross-Entropy Counteracts the Confound Whereas naive estimation creates a bias in the direction of our hypothesis (more frequent verbs exhibit higher valency frame entropy), cross-entropy-based estimation introduces a *conservative* bias *against* our hypothesis: when a verb has very few observations, the observations in the two halves both are very noisy estimates of the actual distribution, increasing cross-entropy. In contrast, for verbs with many observations, both halves have sufficient data to closely estimate the distribution, eliminating the overestimation bias affecting low-frequency verbs.

In order to understand this formally, we consider two verbs, each with k frames and identical distributions $P(F|V = v)$, but where v_1 has N observations, and v_2 has $2N$ observations. In the limit of an infinite corpus, both verbs would have the same values for both VFC and VFE. However, when N is not much larger than k, there is a substantial amount of probability that not all k frames were observed; that is, the diversity of the frames will be underestimated. Such underestimation is more likely to affect the less frequent verb v_1; hence, the entropy of the observed distribution will tend be smaller for v_1 than for v_2. Underestimating both VFE and VFC more strongly for low-frequency verbs, this phenomenon introduces a circularity confound in the direction of our proposed relationship.

The bias *reverses* under cross-entropy estimation: cross-entropy (2) reflects the true entropy (1) plus $D_{KL}(P||P')$. This KL Divergence converges to zero as the number of observations increases, and will tend to be larger for v_1 than for v_2. The bias is thus again stronger for the infrequent verb v_1, but now is an *overestimation* bias, pointing into the direction *opposite* of our proposed relationship.

Cross-entropy estimation controls for the circularity confound in the sense that a systematic pattern in the direction of our hypothesized relationship must reflect a pattern in the underlying true (population-level) VFE values, as the estimation bias points in the opposite direction. Cross-entropy estimation has two limitations: it replaces one bias by another, and it applies only to VFE, not to VFC.

Subsampling As a second way of mitigating the circularity confound, we performed subsampling such that (i) for each verb with frequency above the subsampling threshold Δ, a sample of Δ observations is taken, and (ii) verbs with frequency below Δ are excluded from the analysis.

Considering again verbs v_1, v_2 with the same distribution over frames but different counts N, $2N$, subsampling caps observations of both verbs

at Δ. If both verbs are included (i.e., $N \geq \Delta$), they are on equal footing in estimation of VFE and VFC – the estimation bias is the same for v_1 and v_2.

Subsampling is applicable to both VFE and VFC, and eliminates effects of verb frequency on the estimation bias. However, verbs with insufficient observations need to be excluded, reducing the number of verbs for which correlations of these measures with verb frequency can be tested.

We determine the subsampling threshold Δ for each language by applying a ratio of 0.1 to the maximum frequency of any verb in that language, but limit the absolute value to between a ceiling of 25 and a floor of 10 to avoid wide fluctuations between languages, which would have ranged from 2 in Mbya Guarani to 663 in German.

On average for 79 languages, subsampling filters out 91.6% of all verbs (SD = 4.7%). The number of verbs filtered out ranges from Russian (9083 out of 10410 verbs, 87.3%) to Mbya Guarani (108 out of 113 verbs, 95.6%). We include only languages where more than 10 verbs remain after subsampling, excluding a further four languages[2].

5 Results

We report primarily results with both subsampling and cross-entropy estimation, as these choices lead to the most conservative estimates for testing our hypothesis. The effect of subsampling is examined below.

We quantify the hypothesized relationships using Spearman's rank correlation coefficient (Spearman, 1904). The results show robust correlation between frequency and our valency metrics: when using VFE, we observe at least moderate correlations ($\rho \geq .40$, following Schober et al., 2018) in 66 out of 75 languages and strong to very strong correlations (defined as $\rho \geq .70$) in 18. The remaining 9 languages show only weak correlations[3]. The mean Spearman's ρ is .59, with a standard deviation of .15. Using VFC measure results in similar correlations, marginally stronger than VFE on average across the 75 languages (mean $\rho = .67$, $SD = .14$), of which 71 show at least moderate correlations, and 31 strong to very strong.

We note that our correlation hypothesis takes the form of an overall trend rather than a strict rule. A more frequent verb is not guaranteed nor expected to always have higher diversity of frames. Outliers to this trend among high-frequency verbs illustrate this: in English, verbs like *arrive* (frequency = 237, VFC = 13, VFE = 3.94) have significantly lower VFC and VFE values than others with similar frequency, e.g. *call* (frequency = 238, VFC = 18, VFE = 5.22) and *work* (frequency = 221, VFC = 18, VFE = 5.23).

Fig. 3a shows how the correlation coefficients vary across the languages and in relation to the sample sizes, using VFE. Larger sample sizes, either measured by the number of verb lexemes or by the overall UD corpora size, reduce cross-lingual variance by allowing for better estimations of the entropy and count[4] measures.

Effects of subsampling Our subsampling procedure as described in §3.2 eliminates the circularity confound by fixing the sample size, albeit at the expense of statistical power. Results without subsampling validate our cautious approach, especially for VFC: the mean ρ between VFC and frequency rank is a staggering 0.996 (SD=.004), i.e., almost perfect correlations for all languages. Performing subsampling but then including verbs below the cutoff results in similar numbers (mean ρ =.996, SD = .073 for VFC). These results are arguably a reflection of the circularity confound, eliminated using subsampling.

A visual confirmation of this effect is possible when we plot individual verb lexemes by their frequency ranks on the x-axis and VFE and VFC on y-axes. As seen in Fig. 4, subsampling introduces a flattening effect to the VFE and VFC metrics of verbs with frequency above the subsampling threshold Δ (i.e., verbs to the left of the dotted lines), such that the plot density makes the correlations less visually clear (which can nevertheless be statistically confirmed); in contrast, the clean correlations for the excluded verbs with frequency below Δ (i.e., verbs to the right) reflect how circularity confound would have clouded our analysis without subsampling.

For the VFE measure, cross-entropy estimation alone reverses the circularity confound, even without subsampling (§ 3.2). The stronger correlation (mean ρ =.849, SD = .073) we observe without subsampling helps to confirm that our hypothesis apply to less frequent verbs that were elsewhere excluded. Fig 3b shows cross-linguistic variation

[2] Manx (10), Mbya Guarani (5), Upper Sorbian (10), Welsh (9)

[3] They are Akkadian (.17), Ancient Hebrew (.27), Bambara (.39), Chinese (.39), Classical Chinese (.39), Japanese (.20), Sanskrit (.38), Turkish (.36), and Urdu (.22).

[4] A version of the plot using the VFC measure, which shows a similar pattern, is shown in Appendix, Fig. 7a.

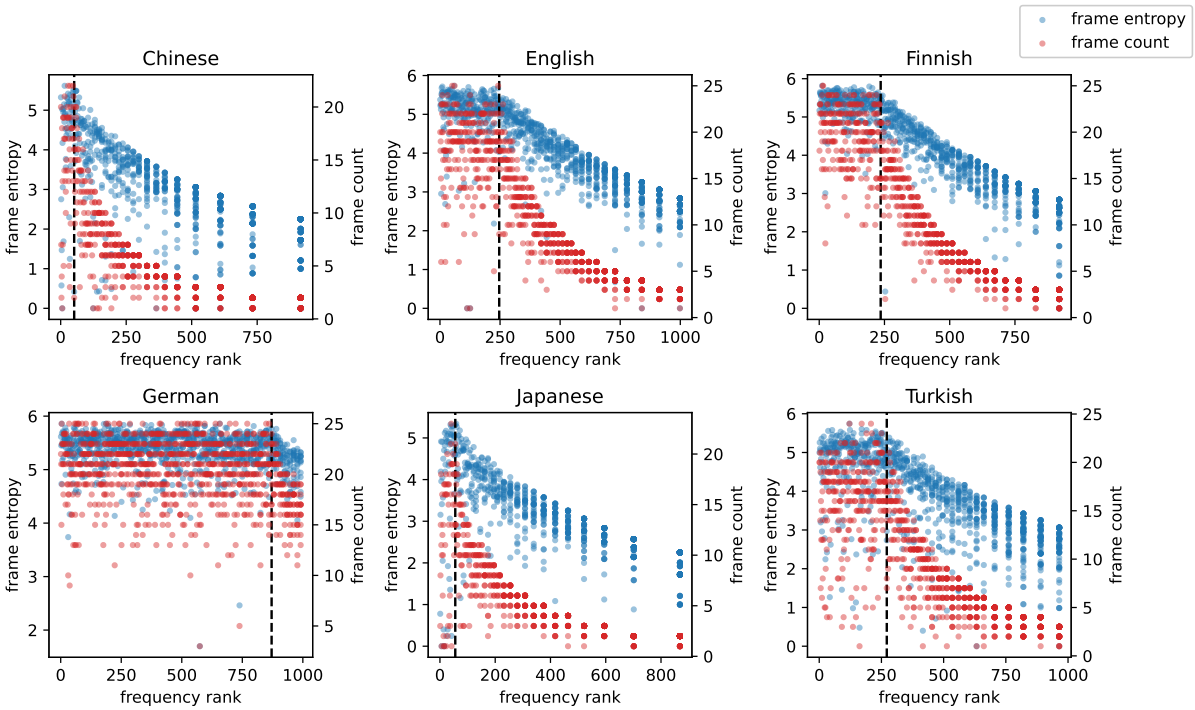


Figure 4: Scatter plots showing verbs with their frequency rank (up to 1000) on the x-axis, VFE and VFC on left and right y-axes. Results use subsampling, the vertical line indicates the subsampling threshold, i.e. only verbs to the left of the line are used in calculating the correlation coefficient in Fig. 3; entropies are estimated as cross-entropies. See Appendix, Fig. 8 for results for all languages, Fig. 9 for results without subsampling.

in the correlation coefficient (VFE) without subsampling. Compared to the subsampled version (Fig 3a), we also observe a greater correlation between the sample size and correlation strength, as the larger number of less frequent verbs for languages with a larger sample size would mean a stronger confound.

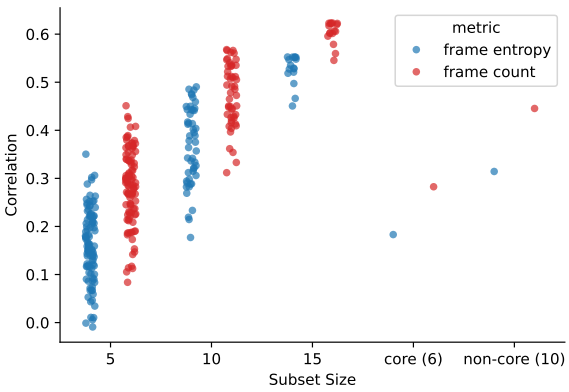


Figure 5: Correlations between frequency rank and VFE / VFC, for the randomly sampled subsets of relations, and core-only and non-core-only relations. Results use subsampling.

Impact of UD Relations We now investigate whether the selection of the UD relations impact the results, i.e., whether certain UD relations drive the observed correlations more than others.

We randomly sampled subsets of sizes 5, 10, 15 from the full set of 16 UD relations, and compared their correlation results with those of a subset including only the 6 core UD relations and of another including only the 10 non-core UD relation. Across the random subsets, larger sets lead to stronger correlations (Figure 5). These results suggest that the observed correlation reflects an across-the-board statistical pattern across a larger set of relation types, and not an artifact of any particular relation type.

Next, in order to compare the effects of the specific relations, we fitted a separate linear mixed-effects model for each relation, predicting the regression coefficient from the presence or absence of that relation in a random subset, with random adjustments for each language. We fitted the model on the set of 1000 randomly sampled subsets of size 5. The models are defined formally in Appendix A.

Fig. 6 shows each relation as a linear function with the fixed effect coefficient as the slope and the fixed effect intercept. The value at UD relation =

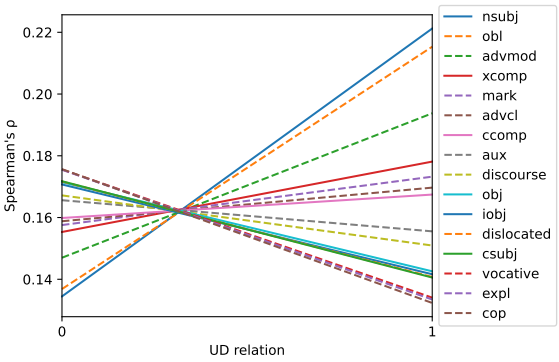


Figure 6: Effects of the 16 relation labels on VFE correlation strength. See Fig. 10 for a version of the plot using VFC. Solid lines indicate core relations; dashed ones indicate non-core relations.

1 is therefore the predicted correlation coefficient when the relation is present. Note that a negative slope in the figure only indicates that including the relation *instead of* other relations on average has a negative effect on correlation strength, i.e., it has a *less positive* effect than average on correlation strength.

Relations differ in their impact. The presence of `nsubj` and `obl` have the strongest positive effect on the correlation strength. The latter is expected, as `obl` encompasses a variety of features often part of frame encoding, including case markings on adjuncts, passive verb use, and dative alternation. The strong positive effect of `nsubj` is more surprising, but the variety of non-predicative uses of verbs, which often involve the absence of or different case markings on `nsubj`, may provide an explanation. There is no evidence that core and non-core relations have systematically different effects based on this metric.

6 Discussion

Results across 79 languages provide strong support to our hypothesis of a positive correlation between a verb's frequency and the diversity of its valency frame. By controlling for the circularity confound, we are able to determine that this correlation stands even after we exclude the effect of the more frequent verbs having more opportunities to appear. This extends previous work studying natural language lexicons and shows that they are optimized for efficiency not just in word-internal proprieties but also in their interface with grammar.

Empirically, our results also complements existing evidence of a correlation between word frequency and semantic ambiguity by testing our hypothesis on the morphosyntax alone (Hoffman et al., 2013). The cross-linguistic coverage further confirms that the results indicate a general trend and are not specific to individual languages.

Our findings show that different syntactic relations differ in their impact on the correlation, but do not support a clear-cut distinction between core and non-core dependents, otherwise understood as arguments and adjuncts. The quantitative trend was supported both by core and non-core dependents, and larger sets of relations strengthened the law. This is evidence in support of a more graded distinction between different verb dependencies rather than a binary one.

While the predicted pattern holds across languages, we observe variation in correlation strength between languages. Understanding whether these differences reflect typologically meaningful variation is an interesting task for future work. One possibility is that the UD annotation scheme does not sufficiently account for cross-linguistic differences in argument encoding: e.g. in Chinese, many outlier verbs that have high frequency but low VFE are *coverbs*, which serve semantic functions similar to prepositions in English while are syntactically verbs (e.g., 自'from', 隨'follow / with'). Their have narrow semantic functions and are associated with a relatively small number of valency frames compared to similarly frequent verbs. This brings into question the strictness of word category boundaries, and fuller picture may need to better situate the verb category within the overall lexicon.

7 Conclusion

Across 79 languages, we have provided evidence for a cross-linguistic quantitative trend in the organization of valency: More frequent verbs are associated with a larger diversity of valency frames, both when estimated by count-based or entropy-based metrics. Crucially, we showed that this pattern is not driven by differences in the number of observations between high- and low-frequency verbs, and persists even when equalizing the number of observations available for each verb. Extending a recent line of research studying the lexicon's efficiency for human communication and language use, these results suggest that such considerations apply not just to word-internal structure, but also to the way words are integrated within the sentence.

Limitations

In §2, we derived our prediction from efficiency considerations in both language production and language comprehension. However, our methods did not allow us to determine which of these pressures are ultimately responsible for the observed pattern. Indeed, it is an ongoing debate to what extent apparently efficient properties of language respect optimization for production or comprehension. It is possible that, in the context of valency organization, these perspectives will make distinct predictions. For instance, theories based on comprehension might induce asymmetries between dependents before and after the head. Teasing these apart in order to determine the contributions of production and comprehension is an interesting problem, beyond our scope here.

A second limitation is that, even though we considered data from 79 languages, we are not able to make truly universal claims. Some language families, such as the Indo-European languages, are substantially over-represented in available corpus data, whereas languages from some parts of the world, in particular Africa and indigenous languages of the Americas, are under-represented, both in the number of languages and language families, as well as in the corpora size.

A third limitation is that corpus data is finite and that in particular the circularity confound forced us to introduce biases to the estimation in order to eliminate that confound. While our results unambiguously show that the predicted correlation exists, a more precise estimation will be affected by these factors.

Fourth, we study a very simple operationalization of valency in terms of UD relations, far simpler than the sophisticated representations typically used in formal linguistic theories. This is due to the fact that such sophisticated representations are not currently available at scale and across languages. While we did verify that the pattern held across languages, we opted for cross-linguistic applicability over language-specific analysis.

References

Peter Ackema. 2015. *Arguments and Adjuncts*, volume 1 of *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)*, pages 246–273. De Gruyter Mouton, Germany.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Collin F. Baker and Arthur Lorenzi. 2020. Exploring Crosslinguistic Frame Alignment. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 77–84, Marseille, France. European Language Resources Association.

Julie E Boland, Michael K Tanenhaus, Susan M Garnsey, and Greg N Carlson. 1995. Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of memory and language*, 34(6):774–806.

John Bowers. 2002. Transitivity. *Linguistic inquiry*, 33(2):183–224.

Joan Bybee. 1998. The emergent lexicon. In *Chicago Linguistic Society*, volume 34, pages 421–435.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, MA.

Noam Chomsky. 1982. *Some concepts and consequences of the theory of government and binding*. MIT press.

Noam Chomsky, Roderick A Jacobs, and Peter S Rosenbaum. 1970. Remarks on nominalization. *1970*, 184:221.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Michael Ellsworth, Collin Baker, and Miriam R. L. Petruck. 2021. FrameNet and Typology. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 61–66, Online. Association for Computational Linguistics.

August Fenk and Gertraud Fenk-Oczlon. 1980. Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß? *Zeitschrift für experimentelle und angewandte Psychologie*, 27:400–414.

Charles J. Fillmore. 1968. The Case for Case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 21–119. Rinehart and Winston, New York.

Charles J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, Seoul, Korea.

Charles J. Fillmore and Collin Baker. 2015. A frames Approach to Semantic Analysis. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*. Oxford University Press.

Adele E. Goldberg. 1992. The inherent semantics of argument structure: The case of the English ditransitive construction. 3(1):37–74.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture Series. University of Chicago Press, Chicago, IL.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in english. *Language*, pages 203–257.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor. 2013. The valency patterns leipzig online database. Max Planck Institute for Evolutionary Anthropology.

Paul Hoffman, Matthew A Lambon Ralph, and Timothy T Rogers. 2013. Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, 45:718–730.

Paul J. Hopper and Joan L. Bybee. 2001. *Frequency and the Emergence of Linguistic Structure*. Typological Studies in Language. John Benjamins Publishing Company, Amsterdam.

J. Kathryn Bock and Richard K. Warren. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1):47–67.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1):109–128.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with Novel Verb Classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Karin Kipper-Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, USA.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Univ. of Chicago Press, Chicago, Ill.

Beth Levin and Malka Rappaport Hovav. 2005. *Argument Realization*. Research Surveys in Linguistics. Cambridge University Press, Cambridge.

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 849–856, Cambridge, MA, USA. MIT Press.

Tal Linzen and T Florian Jaeger. 2015. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive science*, 40(6):1382–1411.

Tal Linzen, Alec Marantz, and Liina Pylkkänen. 2013. Syntactic context effects in visual word recognition: An meg study. *The Mental Lexicon*, 8(2):117–139.

Maryellen MacDonald. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology*, 4.

Kyle Mahowald, Isabelle Dautriche, Mika Braginsky, and Ted Gibson. 2022. Efficient communication and the organization of the lexicon.

Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven T Piantadosi. 2018. Word forms are structured for efficient use. *Cognitive science*, 42(8):3116–3134.

James McCloskey. 1994. *Subjects and subject positions in Irish*. Linguistics Research Center, Cowell College, UCSC.

Shota Momma and Victor S Ferreira. 2019. Beyond linear order: The role of argument structure in speaking. *Cognitive psychology*, 114:101228.

Jesse Mu, Joshua K. Hartshorne, and Timothy O'Donnell. 2017. Evaluating hierarchies of verb argument structure with hierarchical clustering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 986–991. Association for Computational Linguistics.

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

Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Clara Meister, Ethan Gotlieb Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. *arXiv preprint arXiv:2312.03897*.

Tiago Pimentel, Irene Nikkarinen, Kyle Mahowald, Ryan Cotterell, and Damián E. Blasi. 2021. How (non-)optimal is the lexicon? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4426–4438. Association for Computational Linguistics.

Terry Regier, Charles Kemp, and Paul Kay. 2015. Word meanings across languages support efficient communication. *The handbook of language emergence*, pages 237–263.

Patrick Schober, Christa Boer, and Lothar A. Schwarte. 2018. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763.

C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. C. Klincksieck, Paris.

Gertjan van Noord and Gosse Bouma. 1994. Adjuncts and the processing of lexical rules. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological Irregularity Correlates with Frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* The Psycho-Biology of Language: An Introduction to Dynamic Philology. Houghton Mifflin, Oxford, England.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort.* Human Behavior and the Principle of Least Effort. Addison-Wesley Press, Oxford, England.

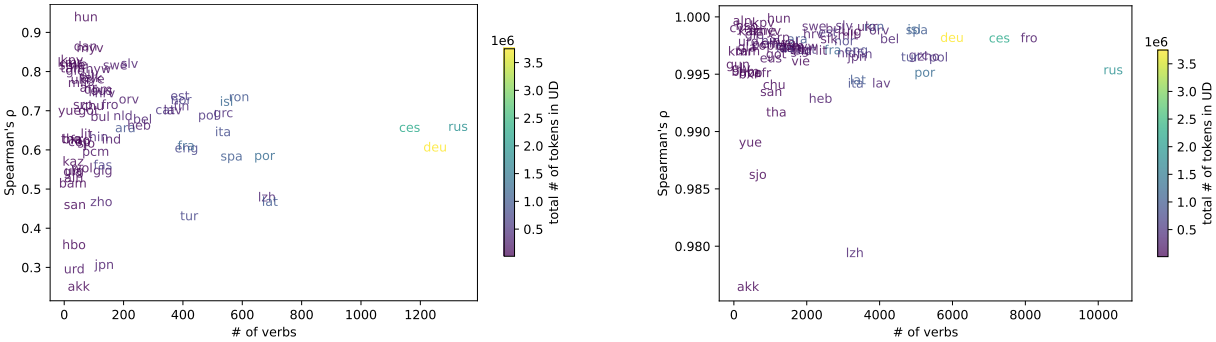
A Linear Mixed-Effect Model

Formally, for the i-th relation across K languages, we fit a linear mixed-effect model with with N different randomly subsampled subsets of relations,

$$\boldsymbol{y} = \alpha^{(i)} + \beta^{(i)} X + Z\boldsymbol{u} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y} \in [-1, 1]^N$ is the outcome variable, i.e. the strength of the rank correlation between verb frequency and valency frame entropy; $X \in \{0, 1\}^N$ indicates, for each of the N subsets, whether the i-th relation is in that subset; $\alpha^{(1)}$ is the fixed effect intercept; $\beta^{(1)}$ is a fixed effect coefficient; Z is a $N \times K$ design matrix of the random effect; u is a $K \times 1$ vector of the random effect of language; and ϵ is a $N \times 1$ column of residuals.

B Additional Figures



(a) results using subsampling

(b) results not using subsampling

Figure 7: Scatter plots showing relationship between Spearman's rank correlation between valency frame count and frequency rank of verbs and the number of verbs across languages. Results use subsampling in subfigure 7a and full data in subfigure 7b; entropies are estimated as cross-entropies. Color shows the total number of tokens for this language in the UD.

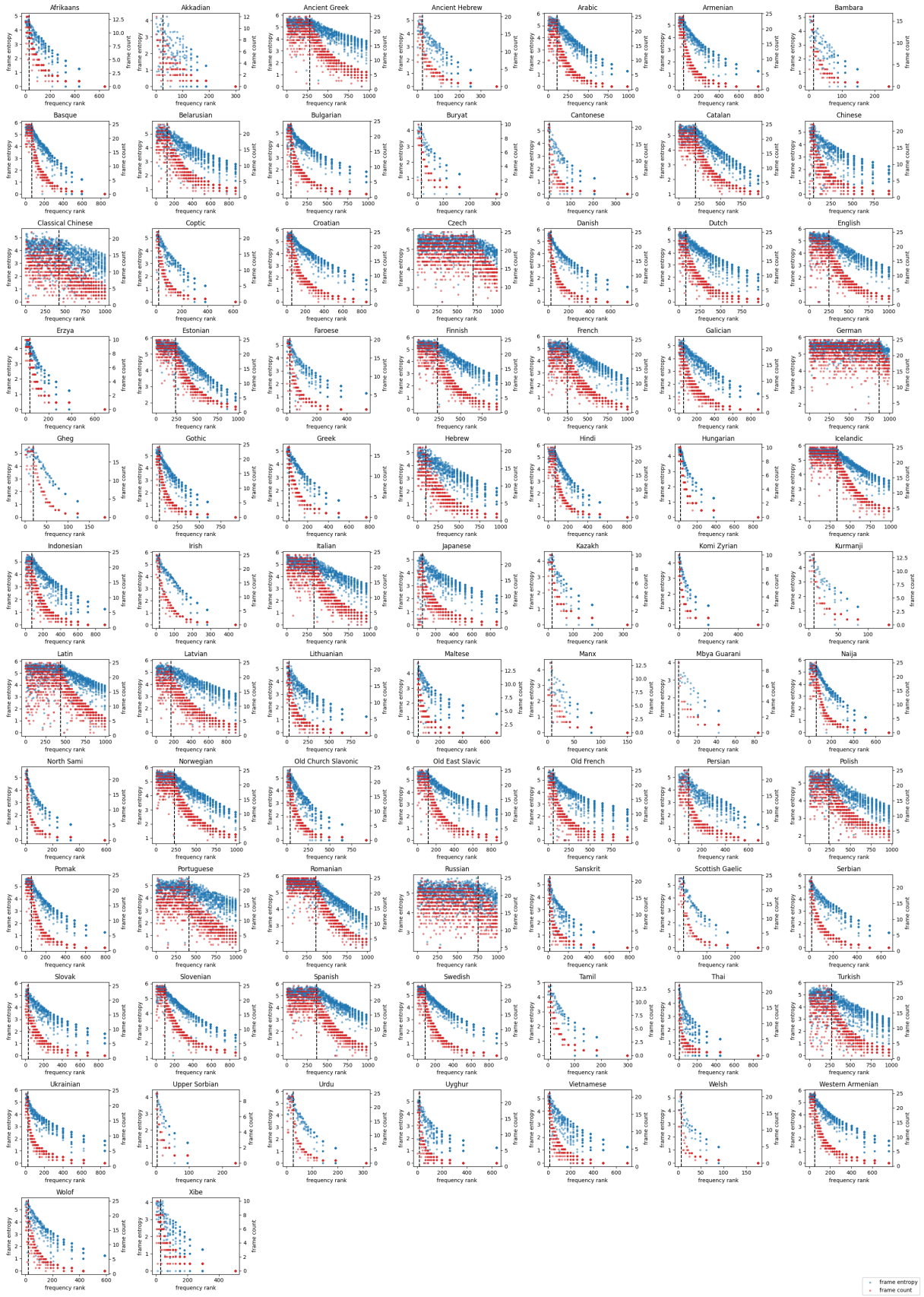


Figure 8: Scatter plots showing relationship between valency frame entropy or frame count and frequency rank of verbs. Results use subsampling, the vertical line showing the subsampling threshold, i.e. only verbs to the left of the line are used in calculating the correlation coefficient in Fig. 1; entropies are estimated as cross-entropies.

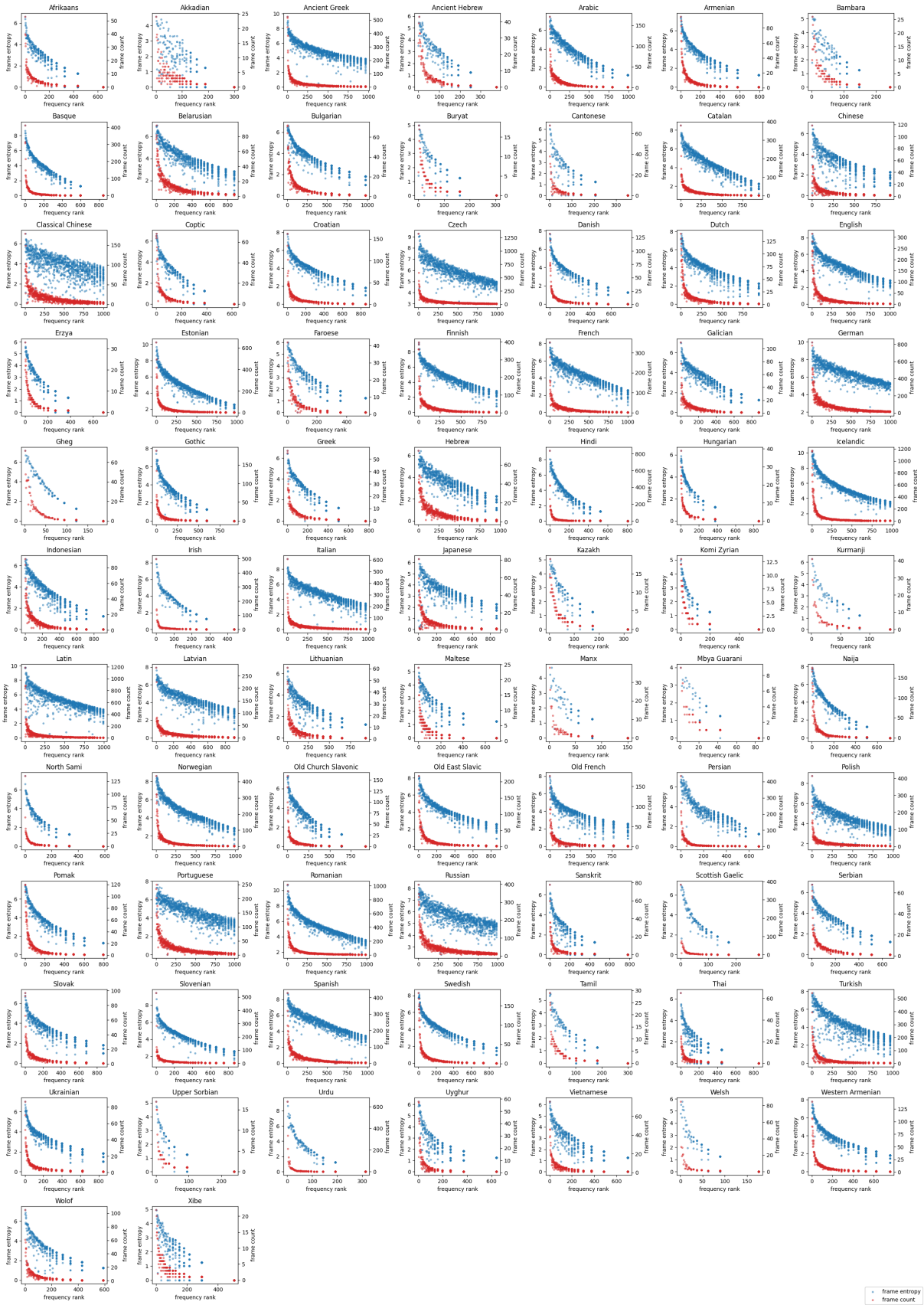


Figure 9: Scatter plots showing relationship between valency frame entropy or frame count and frequency rank of verbs. Results do not use subsampling; entropies are estimated as cross-entropies.

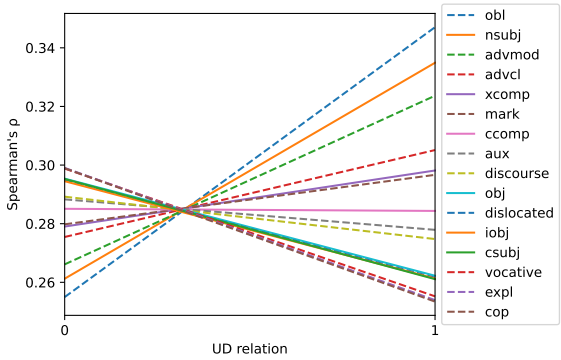


Figure 10: Effects of the inclusion of the 16 relation labels on Spearman's rank correlation coefficient between valency frame count and frequency rank, using VFC metric