The word as a unit of internal predictability

John Mansfield, University of Melbourne

December 2020 preprint; to appear in *Linguistics*

Abstract: A long-standing problem in linguistics is how to define 'word'. Recent research has focused on the incompatibility of diverse definitions, and the challenge of finding a definition that is cross-linguistically applicable (e.g. Haspelmath 2011; Gijn and Zúñiga 2014; Bickel and Zúñiga 2017; Tallman 2020). In this study I take a different approach, asking whether one structure is more word-like than another based on Shannon's (1948) concepts of predictability and information. I hypothesise that word constructions tend to be more 'internally predictable' than phrase constructions, where internal predictability is the degree to which the entropy of one constructional element is reduced by mutual information with another element. I illustrate the method with case studies of complex verbs in German and Murrinhpatha, comparing verbs with selectionally restricted elements against those built from free elements. I propose that this method identifies an important mathematical property of many word-like structures, though I do not expect that it will solve all the problems of wordhood.[1]

Keywords: wordhood, entropy, information theory, morphology, complex verbs

1. Introduction

A great deal of research in linguistics, and the language sciences more generally, relies upon the WORD as a core theoretical concept. Some theories of grammar posit distinct generative systems for words (morphology) and phrases (syntax). Typologists document how the same grammatical function is performed by a word in one language and an affix in another. Corpus linguists and psycholinguists develop theories of language cognition based on statistical relationships among words.

It is therefore rather strange that there should be no widely accepted definition or procedure for identifying words, cross-linguistically or even within a single language (Haspelmath 2011; Tallman 2020a). The word may seem too obvious a concept to deserve this degree of interrogation (Sapir 1921: 33), but there is a serious risk that this 'obvious' concept is actually an artifact of particular writing systems, which have emerged relatively recently, and in relatively few places, given the broader scope of

[1] This paper has benefited from discussion with Christian Döhler, Charles Kemp, William Lane, Frank Mollica, Nicholas Lester, Rachel Nordlinger and Adam Tallman, as well as audiences at the University of Zurich Centre for Linguistics and the University of Melbourne Computational Cognitive Science lab. Further improvements were made thanks to the comments of two anonymous reviewers. The research was funded by the Australian Research Council, grant number DE180100872.

human language (Bloomfield 1933: 178; Saenger 1997; Wray 2015). Meanwhile for speakers of languages without strong orthographic traditions, the distinction between word and phrase may not be obvious at all (Dixon & Aikhenvald 2002: 3, 33), and without reference to clear conventions, there may be little agreement on where exactly the white spaces should appear in writing (Russell 1999; Packard 2000: 14; Rice et al. 2002; Mansfield 2019: 202).

This study proposes a novel, gradient approach to the concept of wordhood. The aim of this approach is to formulate a mathematical measure that will align in interesting ways with the categorical criteria usually referenced to define wordhood. For this I turn to information theory, where 'information' is synonymous with 'entropy', a measure of predictability. My hypothesis is that linguistic construction types with more word-like grammatical properties tend to be more internally predictable, that is to say, one element of the construction is predictive of another. This hypothesis is driven by the intuitive notion of the word as a holistic unit, which suggests that the constituent elements of a complex word may combine more predictably than those of a comparable phrase (Blevins 2016).

An information-theoretic approach to wordhood is motivated by a wave of recent research showing that predictability plays a major role in human language structure and use (for an overview see Gibson et al. 2019). Some of this research uses predictability as a lens to analyse fundamental linguistic properties such as syntactic scope and flexible ordering (Futrell et al. 2019; Lester & Moscoso del Prado Martín in prep.). It is these properties that linguists use to distinguish complex words from phrases, and the current study focuses on another such property, namely selectional restrictiveness (Zwicky & Pullum 1983: 504; Haspelmath 2011: 45). The current study thus fits into a broader research program investigating how categorical properties of linguistic units relate to their gradient, probabilistic properties.

The hypothesised association of wordhood criteria with internal predictabilty can be tested by comparing construction types that are distinguished primarily by wordhood properties. For example, given that fixed ordering is a popular wordhood criterion, one could compare English particle verbs with variable ordering of particle and object (*break up the party ~ break the party up*) against those with fixed order (*get over the incident, *get the incident over*). In this study, I develop the method using construction types that are distinguished primarily by SELECTIONAL RESTRICTIVENESS, where a restricted element is one that appears only in that construction type. German complex verbs have a word-like type with selectionally restrictive preverbs, (e.g. *ver-lassen* 'abandon') and a phrase-like type with unrestricted particles (e.g. *vor#lassen* 'go ahead of'). Using a two-million-word corpus (Menzel 2019), I show that the restricted-preverb construction type is indeed more internally predictable than the particle construction type. A further test on polysynthetic verbs in Murrinhpatha reveals a similar association between selectional restrictiveness on internal predictability. These results support the proposed relationship between wordhood criteria and internal predictability, though the scope of this study is limited to two construction types, and one particular wordhood criterion. Although I propose that internal predictability has an important and interesting association with the concept of the word, I do not expect that it will comprehensively categorise every structure that we may think of as a word. Nor do I think this is possible, given the breadth and ambiguity of the term (§8.2).

An important concept in this study is the CONSTRUCTION TYPE, which is a schematic grammatical structure, rather than a specific linguistic form. My goal is to characterise the wordiness of schematic types such as the German [Particle Verb] structure, and the Murrinhpatha [Classifier-Coverb] structure. These are secondary phenomena, compared to the specific forms such as *verlassen* 'abandon' and *bangkardu* 'see.NFUT.3SG' actually encountered by speakers of German and Murrinhpatha respectively. Linguistic analysis is based on generalisations over specific forms, positing schematic types that share semantic and syntactic properties (Croft 2001). Focusing on these schematic types sets my study apart from a large body of research on the predictability of specific forms in linguistic constructions and collocations (e.g. Ellis & Ferreira–Junior 2009; Gries 2013; Schmid & Küchenhoff 2013).

The remainder of this article is structured as follows. Section §2 outlines the problem of wordhood as established in previous research. Section §3 introduces the concept of entropy and its role in language, while section §4 outlines the proposed method for using entropy to define the internal predictability of construction types. Section §5 introduces German and Murrinhpatha complex verbs, and sections §6 and §7 report internal predictability measures for the two languages. Section §8 summarises the findings and highlights some issues for further research. An appendix addresses issues of entropy measurement in corpus data.

2. The problem with words

There is a long tradition of linguists wrestling with definitions of wordhood (e.g. Sapir 1921; Bloomfield 1933), with several recent studies highlighting the apparent intractability of the problem (Haspelmath 2011; Wray 2015; Ramscar & Port 2016; Bauer 2017 Ch. 2; Tallman 2020a). I here provide a brief synopsis of the standard approach to wordhood, and the problems it faces.

2.1. Morpheme, word and phrase

Most linguistic analysis models human language as hierarchical combinations of symbolic units. The atomic symbolic units, often labelled 'morphemes',[2] are grouped into successively larger constituents. In most linguistic analysis a distinction is drawn between 'word' and 'phrase' constituents, in an asymmetric relationship such that a

[2] I use the term 'morpheme' in the sense of a minimal phonological form associated with semantic content. Haspelmath proposes the alternative term 'morph', since 'morpheme' may also be interpreted (especially within morphology literature) as implying particular theories of grammar and meaning (Haspelmath 2020). I use 'morpheme' in this study as it appears to be the more popular term, and because the definitional issues raised by Haspelmath do not have any immediate bearing on my results.

word may be an element within a phrase, but not vice-versa (Di Sciullo & Williams 1987; Bresnan & Mchombo 1995; Williams 2007). This gives rise to the following hierarchy of constituent types:

(1) morpheme < word < phrase

But if there is a hierarchy of distinct constituent types, how do we identify the cut-off points between levels? Various distributional criteria have been proposed to support level distinctions, and these may be formulated either to distinguish phrases from complex words, or to distinguish words from sub-word constituents of one or more morphemes (i.e. 'bound morphemes'). In this exposition I focus first on the word–morpheme distinction. Many proposed wordhood criteria actually break down into multiple subcriteria and diverse possible interpretations (Tallman 2020a: 7), the complexities of which are beyond the scope of this discussion. I here provide my own interpretation of five distributional criteria distilled from the literature, with idealised examples that appear to show neat word–morpheme distinctions.

i. *Usage as a free-standing utterance*
Words may stand as complete utterances in certain contexts, e.g. *dog* and *walk.* Bound morphemes are infelicitous utterances, e.g. *-ness* and *un-*.

ii. *Selectional restrictiveness*
Words occur in a range of construction types, e.g. *the dog* [Det N], *old dog* [Adj N], *pack of dogs* [N of N], *dog-house* [N-N]. They are selectionally 'promiscuous'. But bound morphemes are selectionally restricted, occuring in only one construction type, e.g. *-ness* only occurs in [Adj-NMLZ].

iii. *Flexible linear position*
The linear position of a word may be influenced by pragmatic effects, e.g. *dogs, I can't deal with*, or free variation, e.g. *finally arrived* ~ *arrived finally*. Bound morphemes are rigidly linearised, e.g. *deaf-ness*, **ness-deaf.*

iv. *Wide scope over coordination*
Words may have wide scope over coordination structures, e.g. *beautiful [dogs and cats]*, but bound morphemes may not, e.g. **un-[conscionable and believable].*

v. *Referentiality*
Words may establish reference, thus facilitating anaphora (e.g. *she has dogs$_i$ and she loves them$_i$*) and modification or quantification (e.g. *I love dogs$_i$ but I don't own any$_i$*). Word-internal morphemes are not available for either anaphora or quantification: **she built a dog$_i$-house as soon as she got them$_i$*, **I have a dog$_i$-house but I don't own any$_i$.*

‘Non-interruptibility’ is another often-cited criterion, but I do not include it here as it is prone to circularity in formulations such as ‘non-interruptibility by another WORD’ (Mugdan 1994: 2552; Haspelmath 2011: 44). Phonological integration and semantic non-compositionality are also often used as diagnostics of wordhood, but in this study I focus primarily on distributional criteria.

The word–phrase distinction can be derived from the word–morpheme distinction: a construction consisting of more than one word element is a phrase, rather than a word. On the other hand, a word may combine with a sub-word constituent to produce a constituent that is still a word, albeit a complex one (e.g. *deaf-ness*, *re-enact*). This introduces a distinction between the MINIMAL WORD (the smallest constituent that satisfies wordhood criteria, whatever they may be) and the RECURSIVE WORD (a constituent made up of a word plus one or more sub-word constituents).[3] The following formulae encapsulate these distinctions, where ‘M’ is a morpheme, $[\_]_{w}$ is a constituent that meets wordhood criteria, and ‘+’ indicates one or more iterations of an element type:

(2) $[\text{M}^{+}]_{w}$ = (Minimal) Word
$[[\text{M}^{+}]_{w}\text{-M}^{+}]_{w}$ = (Recursive) Word
$[[\text{M}^{+}]_{w}\ [\text{M}^{+}]_{w}{}^{+}]_{p}$ = Phrase

Although the designation of multi-word constituents as ‘phrases’ may seem intuitively obvious, it introduces some complications. An element may be designated a ‘word’ due to its properties in one construction type, but the same element may lack these properties when it occurs in a different construction type. This can give rise to purported ‘phrase’ constructions where the constituents fail standard wordhood tests. This is especially evident in lexical compounds, which are generally problematic for the word–phrase distinction (Bloomfield 1933: 180; Haspelmath 2011; Bauer 2017; ten Hacken 2017). Compounds typically combine two or more words, which in the context of the compound fail wordhood tests such as scope over coordination (3) and referentiality (4) (Bauer 2017: 75). But as we will see below, there are other more general problems with wordhood criteria.

(3) a. hot [pies and pastries]
b. *hot[dogs and -sauce] (cf. hotdogs and hotsauce)

(4) a. Our dog$_{i}$ loves the house we built for her$_{i}$.
b. *The dog$_{i}$house is perfect for her$_{i}$.

2.2. Criterial alignment

The most serious problems with wordhood are the lack of consensus on criteria, and the lack of alignment between multiple criteria. Given the lack of a consensual definition,

[3] My use of the term ‘recursive’ refers only to a word element being embedded in a word element. This should not be equated with stricter definitions of syntactic recursion, such as categorial identity across levels, and unlimited levels of embedding (Widmer et al. 2017: 803).

one possible approach is to identify a single criterion, e.g. selectional restrictiveness, and operationalise this as the definition of wordhood. But in this approach 'word' has been replaced by a more explicit concept, and therefore serves no additional analytical purpose. Furthermore, no single criteria seems to consistently reflect intuitions about wordhood (Haspelmath 2011).

The purported word level provides analytical insight if it reflects a non-random alignment or clustering of multiple properties. Some studies, such as Bresnan and Mchombo's (1995) analysis of Bantu classifier constructions, demonstrate alignment of criteria in word vs phrase constructions, but they focus only on specific constructions and do not generalise to an entire language, let alone provide the basis for a cross-linguistic definition. Other studies of multiple criteria have found that they do not align in the way that would be required for a clear categorical word vs phrase distinction (Haspelmath 2011; Gijn & Zúñiga 2014; Bickel & Zúñiga 2017; Tallman 2020a). An example of criterial non-alignment can be seen in languages such as Chintang (Bickel et al. 2007) and Murrinhpatha (Mansfield 2015), which have elements that are selectionally restricted to the inflected verb construction, and cannot stand alone as utterances (and thus appear to be bound morphemes), but exhibit flexible linear order (and thus appear to be words). Grammatical 'function words' often fall between the cracks of the morpheme–word distinction, typically showing more promiscuity than canonical affixes, but on the other hand failing wordhood tests like flexible ordering and isolated utterability. Appeals to an intermediate 'clitic' category don't resolve the problem, as purported clitics also fail criterial alignment tests, and there is no consensus on which criteria are relevant (Spencer & Luis 2012: 220; Haspelmath 2015).[4]

But if we are pessimistic about a formal, categorical distinction between words and phrases, can we nonetheless discover gradient phenomena that converge in support of a fuzzy word concept? One approach is to look for probabilistic associations between criteria. This is the line of investigation pursued by Tallman and colleagues, who systematically test for criterial alignment in languages of the Americas (Tallman et al. 2018; Tallman 2020b). So far, this research has found that many languages do not show criterial alignment beyond chance. I here propose a different approach, testing individual wordhood criteria against a gradient, mathematical measure that is logically independent of distributional properties. For such a measure I turn to the statistical properties of language use, and in particular entropy, a mathematical concept that exhibits striking patterns in language.

3. Entropy and predictability in natural language

Claude Shannon's mathematical theory of information and communication (1948; 1951) would seem, on the face of it, highly relevant to theories of natural language. Yet while

[4] Bruening (2018) raises a different objection against certain lexicalist conceptions of the word–phrase divide, by pointing out that English compounds, purportedly words, can contain phrases as in *that don't-you-dare look* (Bruening 2018: 3).

its impact in electronic engineering and other fields was immediate (Cover & Thomas 2002), its relevance to language remained somewhat neglected during the twentieth century. In the last two decades, however, the relevance of information theory to language has been more widely embraced (Gibson et al. 2019).

A core concept of information theory is ENTROPY, a measure of the predictability of a set of possible outcomes. Higher entropy represents greater unpredictability, which occurs when there are many possible outcomes, none of which is especially probable. The communication of information can be conceived of as an unpredictable outcome: for a message to be informative, it must be unpredictable to some degree. Conversely, when communication is highly predictable, the message is less informative, and it can be compressed, i.e. reduced to fewer symbols in an optimal encoding.

An early application of entropy to natural language was the discovery that symbolic units – words and morphemes – tend to be delimited by highly unpredictable phoneme transitions, compared to the more predictable transitions found within morphemes (e.g. Hafer & Weiss 1974; Brent 1999; following Harris 1955). More recent research has shown that the rate at which information is encoded per second is highly consistent across languages, involving a trade-off of syllable entropy and speech rate (Pellegrino et al. 2011; Coupé et al. 2019). Similarly, there is striking cross-linguistic consistency in the predictability arising from word order (Montemurro & Zanette 2011). Turning to phrasal constituency, it has been shown that words tend to be more interpredictable (i.e. higher mutual information, see §4.1 below) when they are in syntactic dependency relations (Futrell et al. 2019). Similarly, within the English noun phrase, elements closer to the noun are more interpredictable with the noun (Culbertson et al. 2020). These results, among many others, suggest that entropy is a guiding force in the evolution of both phonology and syntax.

Additional support for the relevance of entropy comes from psycholinguistic research showing that the cognitive accessibility of words is linked to the predictability of syntactic contexts. For example, lexical decisions are more quickly reached for words that have a high entropy of surrounding words (McDonald & Shillcock 2001) or immediately preceding words (Baayen 2010). This suggests that words are more strongly represented as independent units if they occur in unpredictable contexts, i.e. their representation is not too strongly associated with some particular context. This is also supported by lexical-decision evidence that high-frequency collocations are processed holistically (Sosa & MacFarlane 2002). There is also evidence that children begin learning by representing phrasal units holistically, and only later breaking them down to facilitate compositionality (e.g. Wray 2002; Bannard & Matthews 2008). The development of compositionality is associated with input entropy: for example, in four-word schematic constructions such as *back in the N*, children can more readily extend the construction to a new N when their previous input involves higher entropy of N (Matthews & Bannard 2010). This suggests that children's acquisition of abstract schematic slots is facilitated by entropy, just as adults' cognitive representation of words is facilitated by entropy of their context.

The syntactic entropy studies cited above are based on orthographic words as their primary data, no doubt because these are the given units that can be extracted from text corpora. But the current study instead uses entropy to interrogate linguistic units, by comparing constituent types that are distinguished according to whether they pass or fail specific wordhood criteria. There has to date been one other study in this direction, in which Geertzen et al. (2016) explore the information content of word boundaries versus morpheme boundaries in English, as well as the comparative information content of word boundaries across English, Estonian, Finnish and Hungarian. They test this using a compression algorithm to measure how much the addition of explicit constituent boundaries adds to the minimal description length of corpus texts. They find that word boundaries add more information than morpheme boundaries in English, and also that English word boundaries are more informative than those of Estonian, Finnish and Hungarian. Geertzen et al. (2016) is an exploratory study that suggests word-like constituents have special informational properties, though it again relies on orthographic words, rather than testing any of the grammatical properties associated with wordhood. Another limitation of the study is that it compares words only against single morphemes, and complete sentences. But the more likely candidates for informationally significant units are constituent types closer to the orthographic word, such as complex stems, or small phrases. The present study aims to further develop the informational analysis of language by focusing on a specific grammatical property, selectional restrictiveness, which is one of the main properties considered to distinguish words from phrasal constructions.

4. Wordhood as internal predictability

The core concept of this study is the internal predictability of a construction type $[X\ Y]_c$, defined as the degree to which the entropy of X is reduced by its mutual information with Y (see details below). My hypothesis is that word-like constructions tend to be more internally predictable than phrase-like constructions. This hypothesis is driven by the intuitive notion of the word as a holistic unit, a *Gestalt*, as well as the psycholinguistic studies mentioned above suggesting that internally predictable constructions are more likely to be represented holistically. This is further supported by studies of corpus co-occurrence frequency, which is closely related to interpredictability, and has been found to correlate with non-compositional semantics and morphological irregularity – both presumably related to holistic storage (Hay 2002; Bybee 2006; Barðdal 2008; O'Donnell 2015 inter alia). If we assume that wordhood is associated with holistic representation, then we should expect word-like construction types to be more internally predictable than phrase-like construction types. I return to this topic in the discussion section (§8.1).

I propose the association of wordhood and internal predictability as a tendency, rather than a rule, since the range of construction types that are word-like in various ways seems too diverse to share any single property. For example some inflectional constructions, which count as words by most criteria, may be quite internally unpredictable.[5] Nonetheless, I expect that internal predictability will prove to be a fruitful formulation for investigating wordhood properties. More specifically, I expect that where two similar construction types are distinguished by a wordhood criterion, the word-like construction type will have greater internal predictability. A corollary of this would be that word-like construction types, all else being equal, have less joint entropy than phrase-like constructions. In information-theoretic terms, words are informationally minimal.

4.1. Construction types, selectional restrictiveness, and entropy

The primary data of this analysis are corpus samples of linguistic construction types. I sample construction types that have two variables over disjoint sets, [X_i Y_j] for 1…i*,* 1…j. These variables are typically represented with word-class or morpheme-class labels such as [Det N], [N-Num], [V-Tns]. The notion of construction type employed here does not assume linear adjacency, but rather includes any compositional structure where the grammar determines the interpretation of the composition. For example the particle verb construction [*break up*] is instantiated in both *break up the party* and *break the party up*, where the the composition of [*break up*] has the same interpretation irrespective of the specific instantiation.

Selectional restrictiveness can be formulated using the intersection of sets occupying slots in distinct construction types. For example, if the German complex verb construction type is [Prev V]$_{cv}$, and the prepositional phrase construction type is [Prep NP]$_{pp}$, then selectionally promiscuous particles like *vor* are those elements in the intersection of Prev and Prep, while selectionally restricted prefixes like *ver-* occur in Prev but not in Prep. Set intersection thus allows us to distinguish constructional subtypes, the particle verb [Part V] and the prefixed verb [Prf-V]. Note that selectional restrictiveness is here defined as a categorical property, which will be compared to the gradient property of internal predictability.

If construction types combine pairs of variables, then the instantiation of these variables with specific morphemes can be expressed probabilistically, e.g. in the [Part V] construction there might be an 0.1 probability that Part=*vor*. Appropriate corpus data will allow us to estimate this probability by counting tokens of [Part V] and calculating what proportion have [*vor* V]. But instead of using a standard probability estimate between 0–1, information theory is based on estimation of SURPRISAL, which uses a negative logarithmic transformation, $S(x) = -\log_2(P(x))$, to express probability on a scale of 0–∞ (Attneave 1959: 6). The most probable or 'expected' outcomes have low surprisal, e.g. $-\log_2(0.95) = 0.07$, and conversely the most improbable or 'surprising' outcomes have high surprisal, e.g. $-\log_2(0.05) = 4.32$, $-\log_2(0.0001) = 13.29$.

[5] In cases of stems combining with agglutinative inflectional affixes, I I expect that internal predictability would be low, as the entropy of affix selection should be fairly independent of the lexical stem. In cases of suppletive allomorphy, there should be very high internal predictability, as the stem is predictive of the affix form. However investigation of inflectional constructions is beyond the scope of this study.

While surprisal measures the unpredictability of a specific outcome, e.g. S(*vor*) = -$\log_2(0.1)$ = 3.32, ENTROPY measures the unpredictability of a set of outcomes. Entropy is defined as the weighted average surprisal of all possible outcomes, where each outcome is weighted by its probability:[6]

$$H(X) = \sum_i P(x_i) * S(x_i)$$

Two important properties of entropy are worth pointing out:

a) All else being equal, variables with a greater number of possible outcomes have greater entropy;

b) Given a variable with *i* possible outcomes, entropy is maximised to the extent that these outcomes are equiprobable.

To make this more concrete, let us take the construction type [Part V], and a hypothetical German corpus sample giving probabilities P(Part) and P(V), with a small set of possible values in each variable. Table 1 illustrates hypothetical frequency counts, and the resulting probability estimates for elements such as P(*vor*)=0.125.

Table 1. Hypothetical counts and probabilities for variables in German [Part V]

	an	aus	vor	über	P(V)
lassen	12	6	4	2	0.375000
gehen	8	2	0	2	0.187500
geben	6	2	0	4	0.187500
stellen	4	6	2	0	0.187500
schieben	1	0	2	0	0.046875
sagen	1	0	0	0	0.015625
P(Part)	0.500	0.250	0.125	0.125	SUM 1.0

Entropy is calculated by transforming these probabilities into surprisals, and calculating the weighted average. For example the entropy of the particle element is given by:

H(Part) = (S(*an*) * P(*an*)) + (S(*aus*) * P(*aus*)) + (S(*vor*) * P(*vor*)) + (S(*über*) * P(*über*))
= (-$\log_2(0.5)$ * 0.5) + (-$\log_2(0.25)$ * 0.25) + (-$\log_2(0.125)$ * 0.125) + (-$\log_2(0.125)$ * 0.125)
= (1 * 0.5) + (2 * 0.25) + (3 * 0.125) + (3 * 0.125)
= 1.75

In the analysis of linguistic constructions, we are most interested in the predictability relationships between variables. We can calculate the SPECIFIC CONDITIONAL ENTROPY of the particle for each verb stem, i.e. H(Part|V) for each value of V, giving conditional

[6] Entropy is often formulated more directly from probability. Given that S(x) = -$\log_2$(P(x)), entropy can therefore be formulated as H(x) = - Σ P(x) * log(P(x)) (e.g. Cover & Thomas 2002: 14).

entropies such as H(Part|*lassen*) = 1.73, and H(Part|*gehen*) = 1.25. These are calculated using the same weighted surprisal summation as above, but using the particle probabilities specific to a certain value of V. For example:

$$\begin{aligned}
H(\text{Part}|\textit{gehen}) &= (S(\textit{an}|\textit{gehen}) * P(\textit{an}|\textit{gehen})) + (S(\textit{aus}|\textit{gehen}) * P(\textit{aus}|\textit{gehen})) \\
&\quad + (S(\textit{vor}|\textit{gehen}) * P(\textit{vor}|\textit{gehen})) + (S(\textit{über}|\textit{gehen}) * P(\textit{über}|\textit{gehen})) \\
&= (-\log_2(0.667) * 0.667) + (-\log_2(0.167) * 0.167) + (-\log_2(0) * 0) + (-\log_2(0.167) * 0.167) \\
&= (0.584 * 0.667) + (2.582 * 0.167) + (0) + (2.582 * 0.167) \\
&= 1.25163
\end{aligned}$$

In our hypothetical data, *gehen* combines with particles more predictably (with lower entropy) than does *lassen*, due to the low surprisal of *an-gehen*. Conditional entropies such as H(Part|*gehen*) tend to be less than the independent entropy H(Part), because the two variables [Part V] are not statistically independent, and knowing the value of V tends to make Part more predictable. The weighted average of these specific conditional entropies gives the GENERAL CONDITIONAL ENTROPY (Cover & Thomas 2002: 17). For example, having calculated the specific conditional entropy for each value of V, we can calculate the general conditional entropy H(Part|V) as follows:

$$\begin{aligned}
H(\text{Part}|V) &= (H(\text{Part}|\textit{lassen}) * P(\textit{lassen})) + (H(\text{Part}|\textit{gehen}) * P(\textit{gehen})) + \\
&\quad (H(\text{Part}|\textit{geben}) * P(\textit{geben})) + (H(\text{Part}|\textit{stellen}) * P(\textit{stellen})) + \\
&\quad (H(\text{Part}|\textit{schieben}) * P(\textit{schieben})) + (H(\text{Part}|\textit{sagen}) * P(\textit{sagen})) \\
&= (1.72957 * 0.375) + (1.25163 * 0.1875) + (1.45915 * 0.1875) \\
&\quad + (1.45915 * 0.1875) + (0.9183 * 0.046875) + (0 * 0.015625) \\
&= 1.473496
\end{aligned}$$

The independent entropy H(Part) = 1.75 has been reduced by some degree to reach the conditional entropy H(Part|V) = 1.47. Closely related to this conditional entropy is the concept of MUTUAL INFORMATION, which is the difference between the independent entropy of one variable and its conditional entropy given a second variable. Thus the mutual information of our hypothetical verb and particle distribution is I(Part;V) = 1.75 – 1.47 = 0.28 (Cover & Thomas 2002: 20).[7] Independent and relational entropies can be nicely summarised by a Venn diagram, as in Figure 1 (cf. Cover & Thomas 2002: 22).

[7] Mutual information is a symmetric relationship between variables, which means that it can be calculated using either variable as a starting point. Thus H(V) = 2.19, and H(V|Part) = 1.91, confirming that I(Part;V) = 2.19 – 1.91 = 0.28 (Cover & Thomas 2002: 21).

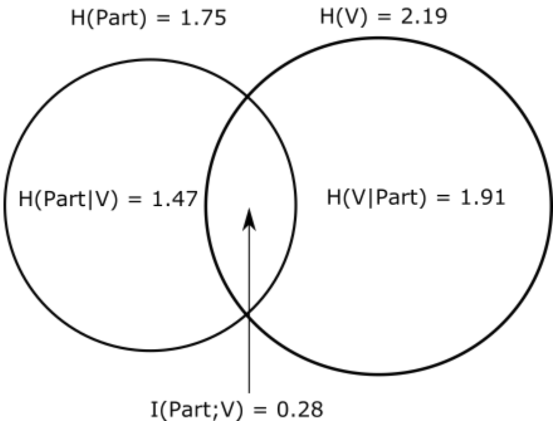

Figure 1. Venn diagram for hypothetical [Part V] independent and relational entropies.

4.2. Internal predictability

In this study, rather than use absolute mutual information I(X;Y), I focus on the proportion of entropy H(X) that is accounted for by I(X;Y). Blevins (2016: 181) calls this 'proportional uncertainty reduction', and with respect to variables in a construction type, I call this the 'internal predictability' (IP) of the construction type $[X\ Y]_C$:

$$IP(X|Y) = I(X;Y)\ /\ H(X)$$

My main hypothesis, then, is that selectional restrictiveness is associated with higher proportional mutual information between elements in a construction type. Put another way, in word-like construction types the information content of the whole (i.e. joint entropy) is much less than the summed information of the parts (i.e. independent entropies). Entropy reduction reflects the intuitive notion that words are informationally minimal.

Using the absolute I(X;Y) value would mean that the magnitude of H(X) in itself would set a limit on the measure, irrespective of the relationship between X and Y. A construction in which H(X) is low can only have a low I(X;Y) measure. I therefore divide I(X;Y) by H(X) to normalise the measure across construction types with varying independent entropy, thus focusing on the relationship between the elements. The importance of this will become clearer in the case studies below.

4.3. Asymmetry and grammatical vs lexical categories

There is inevitably some asymmetry between H(X) and H(Y) in any construction type $[X\ Y]_C$. I focus on the lower-entropy variable, say H(X), because it is here that mutual information accounts for a greater proportion of independent entropy. Taking the lower-entropy H(X), the internal predictability I(X;Y) / H(X) ranges neatly between zero and one. It is possible for the lower-entropy variable to be completely predicted by the higher-entropy variable, but not vice-versa.

In terms of natural languages, H(X) tends to represent the more grammaticalised element of the construction, since grammaticalised elements have fewer possible values than lexical elements, and thus lower entropy.[8] In the example above, particles have lower entropy than verb stems, and I(Part;V) accounts for a greater proportion of H(Part), giving the internal predictability measurement:

IP(Part|V) = 0.28 / 1.75 = 0.16

A complete lack of internal predictability has IP(X|Y) = 0, and we can think of this as an absolute degree of independence or free compositionality between elements in a construction type. This is a limiting case, which does not occur in natural language use (Kilgarriff 2005). Complete internal predictability has IP(X|Y) = 1, where one element is completely dependent on the other, i.e. completely predictable given knowledge of the other. This does occur in natural languages, for example in verbs where a 'thematic' or 'augment' element is consistenly predictable from the root, as in Spanish *camin-a-r* 'walk.INF', or Yankunytjatjara *tju-nku-ku* 'put.FUT' (Goddard 1985: 90). Figure 2 illustrates the Venn diagrams for these limiting cases.

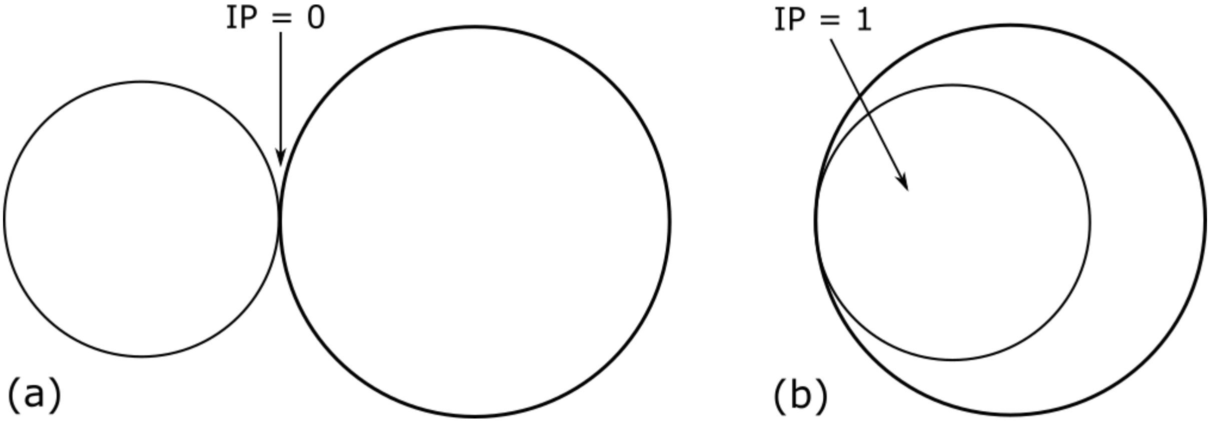

Figure 2. (a) Zero internal predictability, and (b) Complete internal predictability.

5. Words and phrases in complex verbs

To study the relationship between grammatical wordhood criteria and internal predictability, an ideal testing ground is a pair of similar construction types that are distinguished by a standard wordhood criterion. In this study I focus on complex verb constructions distinguished by selectional restrictiveness. Complex verbs are of particular interest since wordhood criteria are often used to distinguish the many varieties of complex verbs found across languages, e.g. preverbs, converbs, polysynthetic verbs,

[8] Note that a grammatical–lexical category distinction is not assumed in this approach, since independent entropy of elements, rather than linguistic analysis, determines the direction of the internal predictability measurement. I follow Brinton and Traugott (2005) in assuming that grammatical–lexical is itself a gradient.

serial verbs and light verbs. Aside from distributional and phonological properties, various degrees of lexicalisation and predictability are often noted in descriptions of complex verb constructions (e.g. McGregor 2002; Aikhenvald 2006; Tersis 2009; Mithun 2020).

German and Murrinhpatha are widely separated by geography, genealogy and typological features, but complex verbs play a central role in the grammar of each. In both languages, a high proportion of verbs combine two distinct classes of elements: a larger lexical class, and a smaller, more grammaticalised class. I label the grammaticalised class 'preverbs' in German and 'finite stems' in Murrinhpatha respectively. Crucially for this study, in both languages complex verb constructions have two main subtypes distinguished primarily by selectional restrictiveness of the grammaticalised element.

German complex verbs can be divided into phrase and word types, according to whether the preverb is a promiscuous element that occurs in other construction types (5), or is selectionally restricted to occur in complex verbs (6). I use the term 'preverb' for the grammaticalised element in both types, and the terms 'particle' and 'prefix' for the promiscuous and selectionally restrictive subtypes (cf. Booij & Van Kemenade 2003; Schultze-Berndt 2003). Particles are promiscuous because they also appear as adverbs or in prepositional phrases (5b). Selectionally restricted prefixes cannot occur in the [Prep NP] construction (6b), or in any other constructional slot. The inflected verb stems are mostly word-like, occurring as simple verbs in a variety of phrasal constructions. Following the formulations above (§2.1), [Part V] is therefore a phrase consisting of two word elements [W W]$_{p}$, while [Prf-V] is a recursive word, consisting of a word preceded by a selectionally restricted morpheme [M-W]$_{w}$.

German complex verb: phrase

(5) a. vor-lassen
in.front-let
[W W]$_{p}$
'let go ahead'

b. vor dem Gebäude
in.front DEF.DAT.NEUT building
'in front of the building'

German complex verb: recursive word

(6) a. ver-lassen
reverse-let
[M-W]$_{w}$
'allow'

b. *ver NP

Murrinhpatha complex verbs can likewise be divided into two main types according to selectional restrictiveness of the grammaticalised element. One type has a finite stem that also occurs in phrasal construction types such as [Subj V] (7b), while the other type has a finite stem that only occurs in complex verb constructions (8b). Thus in both German and Murrinhpatha, the promiscuous elements are disinguished by phrasal constructions in which they occur, though unsurprisingly these are different phrase types in the two languages. There is a major difference from German, however, in that the lexical element, the coverb, is selectionally restricted to occur only in complex verb constructions, and not in other construction types. Thus the promiscuous finite stem type is a recursive word construction [W-M]$_{w}$, while the selectionally restricted finite stem type is a minimal word construction [M-M]$_{w}$.

Murrinhpatha complex verb: recursive word

(7) a. wurran-kath
go.NFUT.3SG-cross.barrier
[W-M]$_{w}$
'(s)he went across (the river, the road etc)'

b. kardu wurran
person go.NFUT.3SG
'the person goes'

Murrinhpatha complex verb: minimal word

(8) a. bam-ngkardu
affect.NFUT.3SG-see
[M-M]$_{w}$
'(s)he sees'

b. *kardu bam
person affect.NFUT.3SG

Both German and Murrinhpatha constructions have been described in some detail elsewhere, and in both cases there has been some discussion of ambiguous wordhood status (Hillert & Ackerman 2002; Müller 2002; van Marle 2002; Eisenberg 2013: 255; Mansfield 2019: 201). Below I briefly review some of the key facts that are relevant for this study.

5.1. German complex verbs

German complex verbs use a few dozen preverbs, combining with a large open class of verb stems.[9] As mentioned above, preverbs can be divided into promiscuous particles that also occur as adverbs or prepositions, and selectionally restricted prefixes that occur only in complex verbs. Verb stems may combine with a range of different preverbs, both particles and prefixes, producing changes of meaning that often relate to aspect (9), argument structure (10) or spatial configuration (11). The meanings of prefixes tend to be

[9] German also has some complex verbs in which the verb stem is preceded by either a noun, e.g. *brust-schwimmen* 'breast stroke', or another verb, e.g. *kennen-lernen* 'get to know' (Eisenberg 2013: 255). However these and any other types are beyond the scope of this study.

more abstract (9b, 10b) and those of particles more concrete (11b) (Los et al. 2012: 176ff.).

(9)	a.	stehen	'stand, be'
	b.	ent-stehen commence-stand	'come into existence'
(10)	a.	arbeiten	'work'
	b.	be-arbeiten affect-work	'work on (something)'
(11)	a.	stoßen	'push, poke'
	b.	durch-stoßen through-push	'push through (something)'

In most German complex verbs the verb stem is a promiscuous element, occurring as a simple verb in various phrasal constructions, e.g. *vor-lassen* 'let go ahead', *lass mich!* 'let me!'. However there are also a minority of verb stems that are selectionally restricted, i.e. do not occur as simple verbs, e.g. *ver-lieren* 'lose', **lieren*.[10] This means that German actually has four complex verb types by selectional restrictiveness, as illustrated in Table 2. In this study we will focus on the more common types, with promiscuous verb stems, while also briefly mentioning the restricted stem types.

Table 2. German complex verb types

	PROMISCUOUS PREVERB	RESTRICTED PREVERB
PROMISCUOUS STEM (common)	*Phrase* [W W]$_{\text{P}}$ e.g. vor-lassen	*Recursive word* [M-W]$_{\text{W}}$ e.g. ver-lassen
RESTRICTED STEM (rare)	*Recursive word* [W-M]$_{\text{W}}$ e.g. aus-statten	*Minimal word* [M-M]$_{\text{W}}$ e.g. ver-lieren

Besides selectional restrictiveness, preverbs are also distinguished by a type of ordering flexibility, in this case conditioned by syntax. Prefixes are rigidly ordered to the left of a verb stem (12), while particles occur to the left of non-finite stems, but with finite stems are instead extraposed at the right edge of the clause (13).

(12) a. Ich muss meinen Hund ver-lassen.
I must my.ACC.MASC dog reverse-let
'I must abandon my dog.'

[10] There are some borderline cases of selectional restrictiveness, depending on whether similar forms in two construction types are treated as being 'the same' or not. In English, for example, analysts may disagree as to whether the forms in [V-*able*] and [*able to* V] are the same, or just etymologically related. The same question applies to some German verb stems, e.g. whether *aus-statten* 'equip' and *an seiner statt* 'instead of him' indicate promiscuous occurrence of *statt* in multiple construction types.

b. Ich ver-lasse meinen Hund.
I reverse-let my.ACC.MASC dog
'I abandon my dog.'

(13) a. Ich muss meinen Hund vor-lassen.
I must my.ACC.MASC dog in.front-let
'I must let my dog go in front.'

b. Ich lasse meinen Hund vor.
I let my.ACC.MASC dog in.front
'I let my dog go in front.'

However, fixed adjacency and selectional restrictiveness do not always align. All prefixes are consistently left-adjacent to verb stems, but some particles exhibit variable adjacency or right-edge extraposition, depending on the lexical stem.[11] Table 3 lists all preverbs extracted in the corpus study below, grouped by selectional restrictiveness and linear flexibility.[12] In this study I focus on selectional restrictiveness, since it provides a clear binary categorisation for each preverb, abstracting away from ordering variation associated with verb stems.

[11] *Hinter* 'behind' appears to be exceptional as a particle that is always left-adjacent to the verb stem.

[12] Preverbs were identified based on the list provided in a German grammar (Dodd et al. 2003: 142ff.). To this list I added several more particles that I found on inspection of my extracted verbs: *bei, fest, fort, frei, hoch* and *los*.

Table 3. German preverbs grouped by wordhood criteria (Dodd et al. 2003: 142ff.)

PROMISCUOUS		SELECTIONALLY RESTRICTED	
Flexible order with all stems		*Rigid order*	
ab	‘from’	be-	‘affect’
an	‘at’	ent- ~ emp-	‘commence’
auf	‘on’	er-	‘complete’
aus	‘out’	ge-	‘terminate’
bei	‘with’	miss-	‘error’
ein†	‘into’	ver-	‘reverse’
entgegen	‘toward’	zer-	‘break’
fern	‘far’		
fest	‘fixed’		
fort	‘away’		
frei	‘free’		
her	‘hither’		
hin	‘toward’		
hoch	‘high’		
los	‘loose’		
mit	‘with’		
nach	‘to’		
vor	‘in front’		
weg	‘away’		
zu	‘toward’		
zurück	‘back’		
zusammen	‘together’		
Rigid order with some or all stems			
durch	‘through’		
hinter	‘behind’		
über	‘over’		
um	‘around’		
unter	‘under’		
voll	‘fully’		
wider	‘against’		
wieder	‘again’		

† *ein* ‘into, in, on’ has somewhat restricted occurrence outside of complex verbs.

There are other systematic differences between the promiscuous and restricted complex verb constructions, such as interruptibility by participial *ge-* and infinitival *zu*, and prosodic differences (cf. Biskup et al. 2011). Another relevant property of the German [Prev V] construction type in general is that it is recursive, i.e. [Prev V] is itself a V, which can be used as the base for a further [Prev V]:

(17) a. $[\text{ein}_{\text{prev}}\text{-}[\text{be}_{\text{prev}}\text{-}[\text{rufen}]_{\text{v}}]_{\text{v}}]_{\text{v}}$
into-affect-call
‘summon (someone)’

b. $[\text{um}_{\text{prev}}\text{-}[\text{her}_{\text{prev}}\text{-}[\text{laufen}]_{\text{v}}]_{\text{v}}]_{\text{v}}$
around-hither-walk
‘stroll around’

This brief account of German complex verbs highlights some of the other wordhood criteria that could provide the focus of future studies. For example one could compare construction types according to criteria of linear flexibility, interruptibility, or prosodic differences. In German verbs these criteria align to a great extent, but of course this need not be the case in other German construction types or in other languages.

5.2. Murrinhpatha complex verbs

Murrinhpatha has 25 different finite stems (not counting reciprocal/reflexive forms),[13] most of which exhibit grammaticalised semantics, and which combine with a large class of 'coverbs', i.e. non-finite lexical stems (Walsh 1976; Street 1987; Blythe 2009; Nordlinger 2015; Mansfield 2019 *inter alia*). Murrinhpatha finite stems are similar to German preverbs, in that they are a small grammaticalised class, and can be divided into a promiscuous type ('free finite stems') and a selectionally restrictive type ('classifier stems', in reference to semantic classification of predicates, see McGregor (2002)). Murrinhpatha verbs potentially host many selectionally restricted elements, including multiple agreement markers and noun incorporation. The whole complex has therefore been characterised as a 'polysynthetic word' (Nordlinger 2017).

An important structural difference between German and Murrinhpatha complex verbs is that German verb inflection is located on the lexical verb stem (e.g. *ver-lasse* PRS.1SG*, ver-lässt* PRS.2SG*, ver-ließ* PST.1SG etc.), while in Murrinhpatha the grammaticalised finite stem carries inflection (e.g. *bam-ngkardu* NFUT.3SG, *dam-ngkardu* NFUT.2SG, *be-ngkardu* PST.3SG). Both inflectional systems use a mixture of affixation and stem ablaut, but in Murrinhpatha there is more suppletive inflection (e.g. *wurran* 'go.NFUT.3SG', *pumpan* 'go.NFUT.3PL', *pa* 'go.IRR.3PAUC') (Mansfield 2016). In this study I treat Murrinhpatha NFUT.3SG forms as citation forms, and omit 'NFUT.3SG' in the glosses.

As with German verb stems, Murrinhpatha coverbs may combine with a range of different grammaticalised elements. Semantic alternations may again relate to aspect and argument structure (18), but Murrinhpatha alternations may also express instruments or body parts (19), and posture or motion of the actor (20) (Nordlinger 2015).

(18) a. mam-ngkan
do-extinguish
'switch off (electric device), put out (fire)'

b. dim-ngkan
sit-extinguish
'It is switched off, extinguished.'

[13] Murrinhpatha has reflexive/reciprocal (RR) versions of its transitive finite stems. The finite stem count is somewhat higher (between 35 and 39) if these are treated as distinct, which is the approach taken in other studies (e.g. Nordlinger 2015; Mansfield 2019). In the corpus data investigated here the transitive/RR pairs generally showed near-identical patterns of lexical composition, which means that treating them as distinct stems would deceptively inflate entropy measures. Another reason for higher finite stem counts in some other work is because variants like *bangam* ~ *bam* and *mangan~mam* have been treated as distinct stems (Mansfield 2019: 114–117).

(19) a. bangam-parl
affect-break
‘smash something (e.g. with hammer)’

b. mungam-parl
use.hands-break
‘snap something by hand’

c. dim-parl
sit.break
‘be broken’

(20) a. dim-rtel
sit-sing
‘sing while sitting’

b. pirrim-rtel
stand-sing
‘sing while standing’

c. wurran-rtel
go-sing
‘sing while moving’

Table 4 lists all the Murrinhpatha finite verb stems, grouped by selectional restrictiveness of finite stems and ordered by corpus frequency (see §5.4 below). Each element is listed in citation form, but it should be kept in mind that this is just one of many inflected forms.

Table 4. Murrinhpatha finite stems grouped by selectional restrictiveness

PROMISCUOUS (‘FREE FINITE STEMS’)		SELECTIONALLY RESTRICTED (‘CLASSIFIERS’)	
mam ~ mangan-†	‘do, say, use hands’††	dam-	‘pierce, use mouth’
kanam	‘be’	bam- ~ bangam-	‘hit’
dim	‘sit’	wurdan-	‘revert’
wurran	‘go’	pan-	‘slash’
nungam	‘travel’	ban-	‘lower’
pirrim	‘stand’	pangam-	‘appear’
yibim	‘lie’	yungan-	‘pull’
kanthin	‘carry’	bim-	‘hear’
pinthim	‘perch’	dirrangan-	‘watch’
		yingam-	‘combine’
		ningam-	‘heat’
		wulam-	‘eat’
		mim-	‘see’
		mungam-	‘use hands’
		dilam-	‘wipe’
		kanthangan-	‘crouch’

† Although *mam* is a promiscuous form, its *mangan-* variant is selectionally restricted to a small number of complex verb forms, and only with the ‘use hands’ meaning.

† † The finite stems *mam ~ mangan* and *dam* are here treated as polysemous. They could alternatively be treated as homophonous but distinct elements (e.g. *mam* ‘do’, *mam* ‘use hands’), but this would introduce the challenge of disambiguation in complex constructions, some of which are semantically opaque.

While most German lexical verb stems are promiscuous, most Murrinhpatha coverbs are selectionally restricted. This means that my study of Murrinhpatha will focus on the contrast between recursive words and minimal words, whereas in German the main contrast is between phrases and recursive words. However Murrinhpatha does also have a few promiscuous coverbs: for example *matharr* 'sick' can appear in the complex verb *wurran-matharr* 'go around sick', or the copula construction *kardu matharr* 'the person is sick'. Therefore Murrinhpatha has the same four complex verb types as German by selectional restrictiveness, as illustrated in Table 5. But here it is the restricted coverb types that are common, and therefore will be the focus of the study. Promiscuous coverb constructions are quite rare, representing only 4% of coverb types and 9% of tokens in our corpus data (details in §5.4 and §7 below), which is insufficient to facilitate entropy analysis.

Table 5. Murrinhpatha complex verb types

	PROMISCUOUS FINITE STEM	RESTRICTED CLASSIFIER
PROMISCUOUS COVERB (rare)	*Phrase* [W W]$_{P}$ e.g. wurran-matharr	*Recursive word* [M-W]$_{W}$ e.g. dirrangan-birl
RESTRICTED COVERB (common)	*Recursive word* [W-M]$_{W}$ e.g. wurran-kath	*Minimal word* [M-M]$_{W}$ e.g. bam-ngkardu

5.3. Selectional restrictiveness and internal predictability

My main hypothesis is that construction subtypes differentiated by a wordhood criterion should show a difference in internal predictability. I therefore expect that the German recursive word subtype [M-W]$_{W}$ will be more internally predictable than the phrase subtype [W W]$_{P}$. Assuming that the effect of selectionally restrictive elements is cumulative, I furthermore expect that in Murrinhpatha the minimal word type [M-M]$_{W}$ will be more internally predictable than the recursive word type [W-M]$_{W}$.

Internal predictability is not an inevitable or trivial property of selectional restrictiveness. It is an empirical question, stimulated by the intuition that words are more holistic units than phrases (§4). Importantly, the question asked here is NOT whether selectionally restricted elements have lower entropy than promiscuous elements. This will almost certainly be the case for German preverbs, simply because there are many more promiscuous particles than restricted prefixes. Nor does internal predictability depend upon entropies beyond a particular construction type. For example, the aggregate entropy of German particles both in complex verbs and in prepositional phrases would make the greater predictability of selectionally restrictive elements all but inevitable (for further discussion see §8.1). The question asked in this study is to what degree one element of a construction type makes the other more predictable, focusing only on distributions within that construction type. This proportional reduction is not intrinsically determined by the independent entropy of either element, or by their distributions in other construction types.

5.4. Corpus data and method

German corpus data was extracted from the manually annotated section of the Hamburg Dependency Treebank (Menzel 2019), consisting of some two million words of written Standard German, sourced from a German news website with a technology focus. Sourcing from a single news website carries a risk that the text may be somewhat semantically homogeneous, and this may be improved upon in future research using more diverse corpus samples. On the other hand, the main advantage of the treebank corpus is that its syntactic annotation allows for both conjoined and separated complex verbs to be accurately extracted (23). [Prev V] citation forms were extracted, ignoring inflectional variants (24). As noted above, there are some recursive instances of German [Prev V], each of which yielded multiple tokens (25). Altogether the corpus yielded 77,946 tokens of German complex verbs.[14]

		Corpus source phrase	*Citation form(s) extracted*
(23)	a.	Microsoft […] feierte den Sieg für 'Microsoft und für alle, die sich für Innovation in der High-Tech-Industrie einsetzen.'	ein-setzen
	b.	Um das eigene Programm mit AIM kompatibel zu machen, setzte Microsoft die Methode des reverse engineering ein.	ein-setzen
(24)	a.	Man würde […] eine Verschlechterung gegenüber dem aktuellen Produkt erwarten.	er-warten
	b.	Anrufer erwartete bei der Telekom […] öfter das Besetzt-Zeichen.	er-warten
(25)		Eine Cindy Crawford etwa ist mit 80000 Dollar veranschlagt.	an-schlagen, ver-anschlagen

Selectional restrictiveness of preverbs was determined with reference to a German grammar (Dodd et al. 2003: 142ff.), while restrictiveness of verb stems was determined by checking whether they appear in a wordlist from the seven billion-word DeReKo corpus (Belica et al. 2014).

[14] Scripts and raw data used in this study are available at: https://github.com/jbmansfield/Word-predictability

Murrinhpatha complex verbs were extracted from a corpus of around 100,000 words of morphologically annotated Murrinhpatha speech, comprising a range of genres (Mansfield et al. 2020). This yielded 8203 complex verb tokens in total, but this was filtered down to 6041 tokens, since the remaining tokens were produced in elicitation tasks that do not reflect naturalistic discourse patterns. The difference in token counts between German and Murrinhpatha has potential implications for entropy measurements, which I address in the appendix.

Murrinhpatha [Fin Cov] citation forms were extracted for each example, abstracting away from inflectional elements (26) as well as other components of the polysynthetic verb complex (27).

		Corpus source verb complex	*Citation form(s) extracted*
(26)	a.	dam-winhipak pierce.3SG.NFUT-pour ‘(s)he poured it’	dam-winhipak
	b.	tha-winhipak pierce.2SG.IRR-pour ‘you pour it!’	dam-winhipak
	c.	dem-winhipak-dim pierce.RR.3SG.NFUT-pour-sit.IPFV ‘(s)he is pouring it for herself/himself’	dam-winhipak
(27)	a.	me-ngintha-thap-thap-tha-kardi use.hand.3SG.PST-DU.F-touch-touch-PST-be.IPFV ‘two women were touching it’	mam-thap
	b.	pumem-ngka-thap-thap-ngime use.hand.RR.3PL.NFUT-face-touch-touch-PAUC.F ‘the women are touching each others’ faces’	mam-thap

Selectional restrictiveness of Murrinhpatha finite stems and coverbs was determined based on the author’s own Murrinhpatha fieldwork materials and the Murrinhpatha dictionary (e.g. Street 2012).

Both corpus samples have their limitations: for German, the homogeneous subject matter; for Murrinhpatha, the modest word count. However these limitations do not introduce any obvious confounds into the comparison of promiscuous and restricted construction types, and thus these provide appropriate data sources for a study of internal predictability and wordhood.

6. Results: Internal predictability of German complex verbs

In this section I compare the internal predictability of the two main German complex verb construction types: those with free particles and those with prefixes. By the criterion of selectional restrictiveness, these are phrase and recursive word types respectively. The two types have approximately equal token frequency in the German corpus ([Part V] = 37,700; [Prf-V] = 40,246).

6.1. Particle versus prefix verbs

To investigate the difference between construction types with particles and prefixes, I first exclude verbs that have selectionally restricted stems (10% of the total), as these would introduce another dimension of variance into the dataset. I report briefly on the smaller restricted-stem dataset following the main analysis.

The proportional corpus frequencies of preverbs within each construction type, interpreted as probability estimates, are shown in Figure 3.[15] These probability distributions underpin the independent entropies H(Part) and H(Prf) in the two construction types. There is a larger set of particles (N=30), with many rare elements, producing an independent entropy of H(Part) = 3.94. Prefixes are a smaller set (N=7), and most of the probability density is accounted for by three frequent prefixes: *be-*, *ver-* and *er-*, producing a much lower entropy H(Prf) = 2.10.

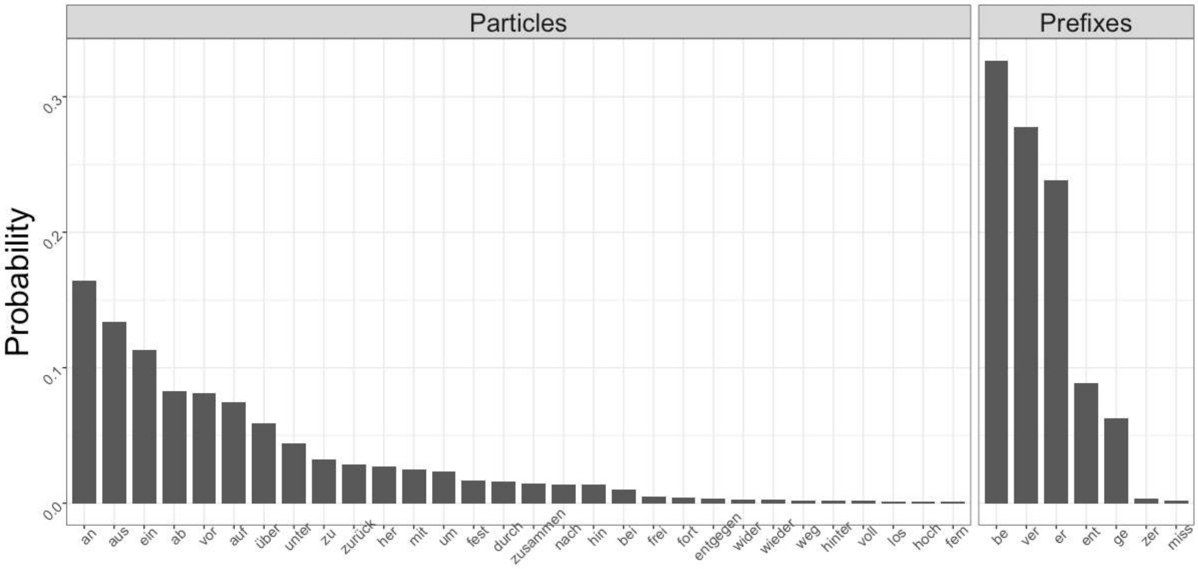

Figure 3. Probability distributions of German verbal particles and prefixes.

In the phrasal construction type, the entropy of free particles H(Part) is only somewhat reduced in the context of specific verb stems. Table 7 provides illustrative examples of the probabilities with which verb stems combine with particles. Each stem is also given an internal predictability measure, similar to that outlined in §4, though the figures here

[15] See script extract-compounds.R

are for specific verb stems rather than the weighted average of all stems.[16] Internal predictability approaches 1 as particle selection by verb stem becomes more predictable. For each of the verb stems shown here, internal predictability is in the middle of the range, as the conditional particle entropies are around half the independent particle entropy, H(Part) = 3.94.

Table 7. Typical particle verbs, with intermediate proportional predictability.

VERB STEM	*arbeiten*	P(Part)	*streichen*	P(Part)	*werfen*	P(Part)
PARTICLES	*zusammen*	0.58	*unter*	0.72	*vor*	0.76
	aus	0.12	*ein*	0.13	*unter*	0.06
	über	0.10	*zusammen*	0.09	*auf*	0.05
	mit	0.08	*durch*	0.02	*ab*	0.01
	ein	0.04	*um*	0.02	*über*	0.01
	ab	0.03	*zurück*	0.02	*weg*	< 0.01
	vor	0.01			*um*	< 0.01
	durch	0.01			*aus*	< 0.01
	zu	0.01			*an*	< 0.01
	auf	< 0.01			*ein*	< 0.01
	hin	< 0.01			*hin*	< 0.01
	um	< 0.01				
Conditional entropy $H(Part \mid V)$		2.18		1.60		1.46
Internal predictability $IP(Part \mid V)$		0.45		0.60		0.63

In the recursive-word construction type, the entropy of prefixes is more substantially reduced by verb stems. Table 8 provides illustrative examples of verb stems and prefixes. Internal predictability measures are close to 1, because conditional prefix entropies account for a small proportion prefixes' independent entropy, H(Prf) = 2.10.

Table 8. Typical prefix verbs, with high proportional predictability.

VERB STEM	*kaufen*	P(Prf)	*kommen*	P(Prf)	*reichen*	P(Prf)
PREFIXES	*ver*	0.99	*be*	0.99	*er*	1.00
	er	0.01	*ent*	0.01		
			ver	< 0.01		
Conditional entropy $H(Prf \mid V)$		0.07		0.12		0.00
Internal predictability $IP(Prf \mid V)$		0.97		0.94		1.00

The overall internal predictability measurement for each construction type is formulated as the weighted average of conditional entropies for all verb stems (§4.2). Figure 4 illustrates the distribution of internal predictability for verb stems in each construction type – though this does not show the probability of each verb stem, which is the weighting factor used to produce the overall measure. The y-axis counts verb stem types, which are grouped on the x-axis by internal predictability. Only stems with token count

[16] See script lex-gram-ent.R.

≥ 10 are included in this stem histogram and those that follow, as entropy measures become less reliable with small counts (see appendix).

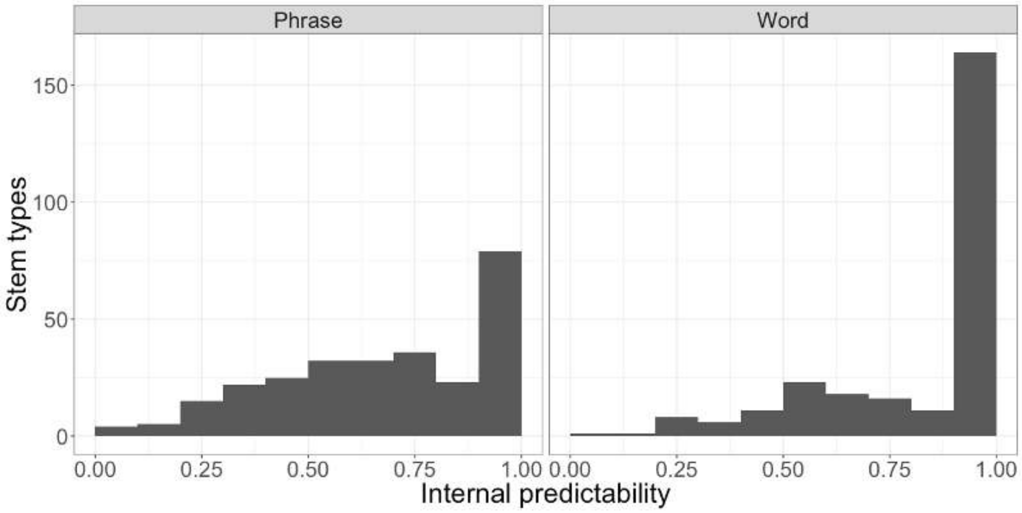

Figure 4. German preverb predictability, in phrase and word type constructions.

In both construction types, the distribution appears to fall into two groups, one with a normal distribution and another approaching the upper limit of 1.0. Yet there is also an important difference in the distributions. In phrasal verbs, the majority of verb stems (71%) are in a group with a right-skewed normal distribution between 0 and 0.9, and a minority (29%) are in the upper-limit group, i.e. above 0.9. But in word-type verbs, the minority of stems (37%) are in the normally distributed group, and the majority (63%) in the upper-limit group. Thus the example stems shown in Tables 7 and 8 above are representative of the main tendency in each group.

These distributions supports the hypothesis that selectional restrictiveness is associated with greater predictability. Median preverb predictability in the phrasal group is 0.71, and the median in the word group is 1.0. Since the overall distributions are not normal, the non-parametric Wilcoxon rank sum was used to test for a statistical difference between the two groups, finding that they are extremely unlikely to be drawn from the same underlying distribution ($p < 0.001$, effect size $r = 0.34$).

Figure 5 shows the overall entropy, conditional entropy and mutual information of the construction types, i.e. the weighted average figures for all stem elements, including the low-count stems (N<10) not shown in the histograms above.[17] The independent entropy of stems is very similar in the two types (7.12 vs 7.21), while preverb entropy is much higher in the phrasal type (3.94 vs 2.10), as we saw in Figure 1 above. Absolute mutual information is actually higher in the phrasal type (2.04 vs 1.60), but it accounts for a greater PROPORTION of preverb entropy in the word type. This demonstrates the importance of normalising I(X;Y) measures against H(X). The phrasal

[17] Low-count stems have less reliable conditional entropy estimates, but they can be included in the overall weighted average conditional entropy since their low probability correspondingly reduces their contribution to the weighted average. For further discussion see the appendix.

type has significantly higher H(X), and we factor this out to focus on the relationship between elements. Using this proportional measure, the phrasal construction type has as mid-range internal predictability score, IP(Part|V) = 2.04 / 3.94 = 0.52, while the word construction type has a higher IP(Prev|V) = 1.60 / 2.10 = 0.76. Given that the IP measure ranges from 0 to 1, this difference of 0.24 can be regarded as substantial.

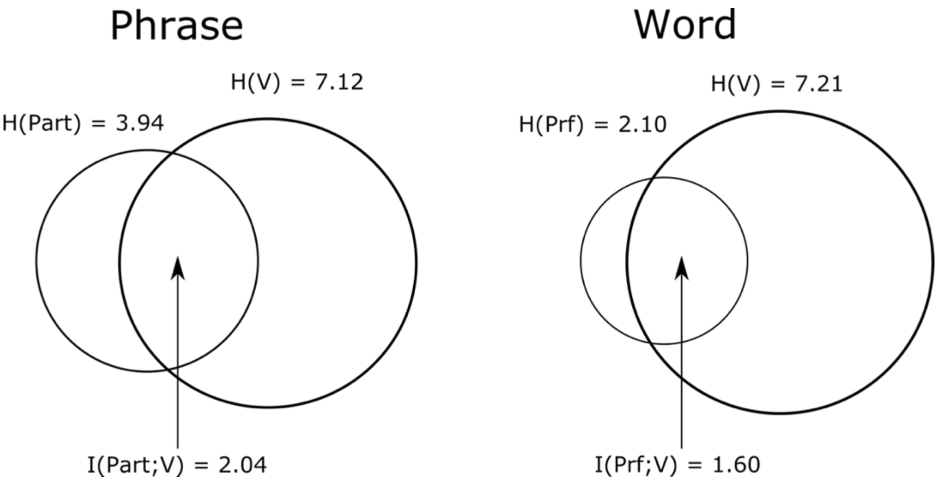

Figure 5. Independent entropies and mutual information for German: phrase type (particle) verbs and word type (prefix) verbs.

6.2. The role of stems with low frequency as simple verbs

We saw above that more verb stems in the phrasal construction type have moderate preverb predictability, while more in the word construction type have almost complete preverb predictability. But we also saw that the phrasal verb construction does have a substantial group of stems with high preverb predictability, albeit not as many as the word type. Investigation of this group shows that the token frequency of stems as simple verbs explains this to some extent.

Figure 6 repeats the histogram above, but now shaded to distinguish verb stems with higher simple-verb token frequency (N ≥ 20) in dark grey, and lower frequency (N < 20) stacked on top in light grey.[18] These simple verb frequencies are extracted from the same corpus sample as the complex verbs, and the N=20 threshold was determined heuristically by plotting frequency against predictability. In the phrasal type (left panel), we can see that stems with higher simple-verb frequency (dark grey) have a mostly normal distribution. Verb stems with lower simple-verb frequency (light grey) instead tend towards the upper limit of preverb predictability. In the word type (right panel), verb stems in general mostly have very high preverb predictability, irrespective of their simple-verb frequency.

[18] See script explore-factors.R

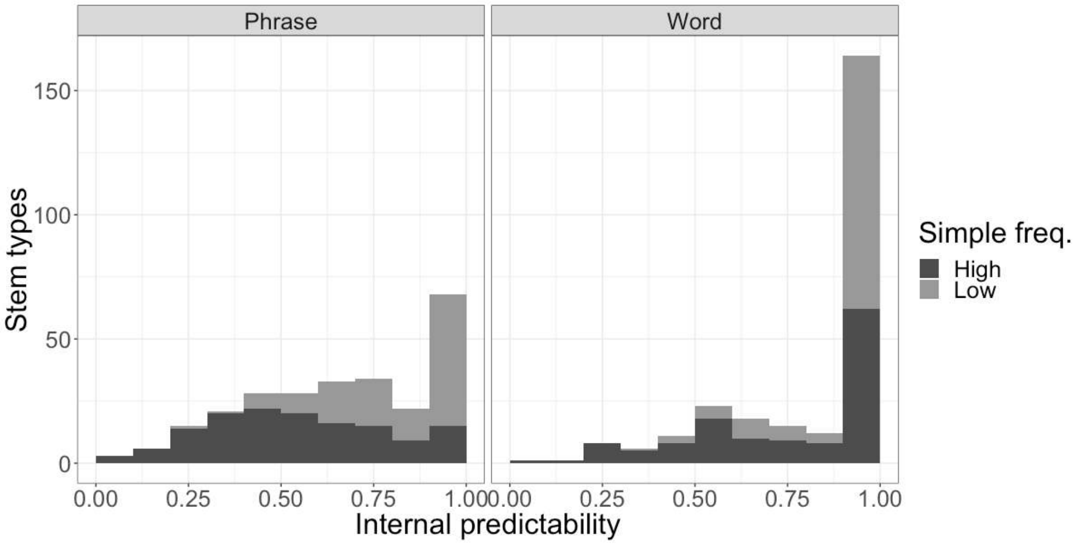

Figure 6. German verb stems internal predictability values, in phrase and word type constructions.

Table 9 provides examples of stems that have lower simple verb frequency, and high internal predictability in the phrasal verb construction. These are all stems that occur more frequently in complex verbs than as simple verbs in the corpus sample.

Table 9. Atypical particle verb stems, with low simple-verb frequency and high particle predictability.

VERB STEM		*sacken*		*tauchen*		*weiten*
SIMPLE N		12		1		0
COMPLEX N		32		106		67
PARTICLE PROBS.	*ab*	1.00	*auf*	0.95	*aus*	1.00
			ein	0.05		
Conditional entropy H(Part|V)		0.00		0.29		0.00
IP(Part|V)		1.00		0.92		1.00

These results show that there is an additional factor at play in the internal predictability of German complex verbs. Selectional restrictiveness of the preverb has an influence, but the frequency of the stem as simple verb also has an influence, at least in particle verbs. This suggests that the internal predictability of a construction type is influenced not just by the selectional restrictiveness of its elements, but also by the frequency with which promiscuous elements appear in the other construction types. In the discussion section below I discuss how this might be further explored in models of cognitive representation (§8.1).

6.3. Complex verbs with selectionally restricted stems

As noted above, a minority of German complex verbs have selectionally restricted stems (7760 tokens, 10% of total). These were set aside from the analysis of particles vs prefixes, since restricted stems would introduce an unwanted extra dimension of variance. A large majority of restricted-stem verbs have prefixes (forming minimal words), rather than particles (forming recursive words); but restricted-stem verbs have very high preverb predictability, irrespective of preverb selectional restrictiveness. This is in line with our expectation that restricted stems, like any selectionally restricted element, should be associated with internal predictability.

Figure 7 illustrates type counts of restricted stems on the y-axis grouped by preverb predictability on the x-axis. The recursive word (particle) type on the left has only fourteen stems that meet the N ≥ 10 token threshold, and all of these have highly predictable particle selection. The minimal word (prefix) type on the right provides more data, with 80 stems, and here the vast majority have highly predictable prefix selection. The overall internal predictability of the minimal word construction type is IP(Prf|Stem) = 0.97.

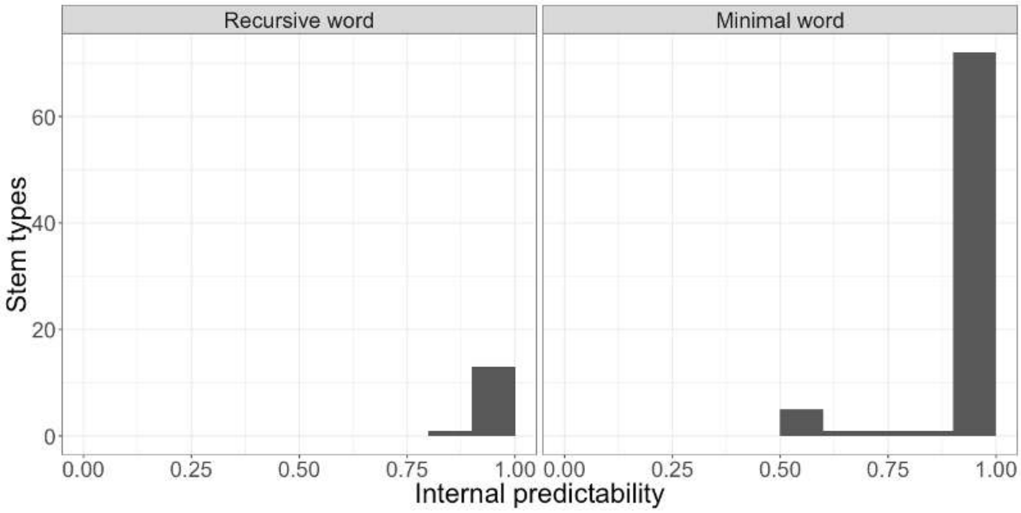

Figure 7. German complex verb internal predictability with bound stems.

7. Results: Internal predictability of Murrinhpatha complex verbs

We turn now to Murrinhpatha, where again we compare verbs with promiscuous grammaticalised elements versus those with selectionally restricted grammaticalised elements. Recall that in this case the grammaticalised element carries inflection, and thus is termed the ‘finite stem’. Furthermore, unlike the mostly promiscuous verb stems of German, Murrinhpatha’s main lexical element, the ‘coverb’, is usually a selectionally restricted element. Thus the comparison here is between a recursive word construction type with (promiscuous) free finite stems [W-M]$_{\text{W}}$, and a minimal word construction type with (selectionally restricted) classifier stems [M-M]$_{\text{W}}$. The corpus has fairly similar token counts of recursive-word verbs (N=2834) and minimal-word verbs (N=3207).

Figure 8 illustrates the probability distributions of Murrinhpatha finite stems, showing that promiscuous and restricted finite stems have similar distribution profiles. Whereas in German the selectionally restricted prefixes are much fewer in number than promiscuous particles, in Murrinhpatha there are somewhat more classifier stems

(N=15) than free finite stems (N=9).[19] The finite stem entropies produced by these distributions are fairly similar, for free finite stems H(Ffs) = 2.45, and for classifier stems H(Cls) = 2.82.

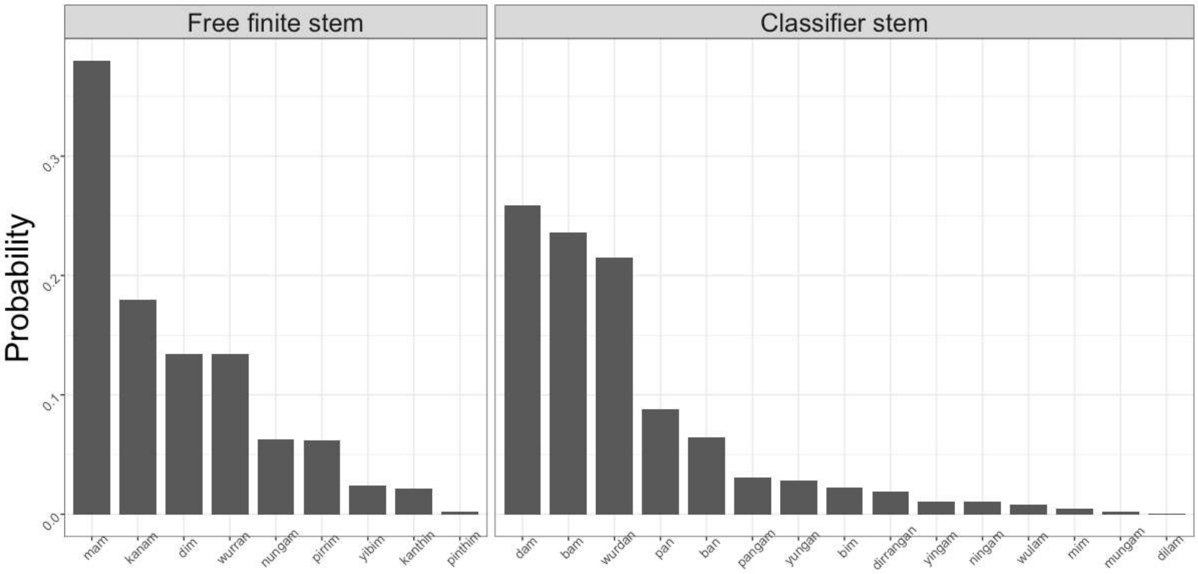

Figure 8. Probability distributions of Murrinhpatha free finite stems (left) and bound classifiers (right).

7.1. Promiscuous versus selectionally restricted finite stems

Tables 10 and 11 provide illustrative examples of coverbs with free finite stems (recursive-words) and classifier stems (minimal-words) respectively. Coverbs combining with free finite stems have mid-range internal predictability, since the conditional entropy of finite stems accounts for a substantial proportion of their independent entropy, H(Ffs) = 2.45. Coverbs combining with classifier stems have high internal predictability, since conditional entropy is only a small proportion of their independent entropy, H(Cls) = 2.82.

Table 10. Typical recursive-word coverbs, with mid-range internal predictability.

COVERB	*-kut*	P(Ffs)	*-mardawith*	P(Ffs)	*-pup*	P(Ffs)
PROMISCUOUS FINITE STEM	*kanam*	0.64	*dim*	0.64	*yibim*	0.46
	mam	0.13	*kanthin*	0.16	*kanam*	0.44
	dim	0.10	*wurran*	0.14	*wurran*	0.05
	wurran	0.10	*pirrim*	0.04	*dim*	0.03
			nungam	0.01	*pirrim*	0.03
Conditional entropy H(Ffs|C)		1.68		1.55		1.72
Internal predictability IP(Ffs|C)		0.31		0.37		0.30

[19] One classifier stem, *kanthangan-* ‘crouch’, is attested in older documentary materials (Street 1987) but does not appear in our corpus.

Table 11. Typical minimal-word coverbs, with high finite-stem predictability.

COVERB	*-ruy*	P(Cls)	*-pirnturt*	P(Cls)	*-wuy*	P(Cls)
CLASSIFIER STEM	*pangam-*	0.96	*dam-*	1.00	yungam	0.97
	dam-	0.04			bangam	0.03
Conditional entropy H(Cls|C)		0.24		0.00		0.19
Internal predictability IP(Cls|C)		0.91		1.00		0.93

Figure 9 illustrates the distribution of internal predictability for coverbs in each construction type. This shows a fairly similar pattern to German construction types with promiscuous and restricted preverbs. Again each construction type has a mixture of coverbs with intermediate finite stem predictability, and others approaching the upper limit of predictability. In verbs with free finite stems the groups are about the same size, while in verbs with classifier stems the highly predictable group is much larger. Again the Wilcoxon rank sum test shows that the two groups are extremely unlikely to be drawn from the same underlying distribution ($p < 0.001$, effect size $r = 0.37$).

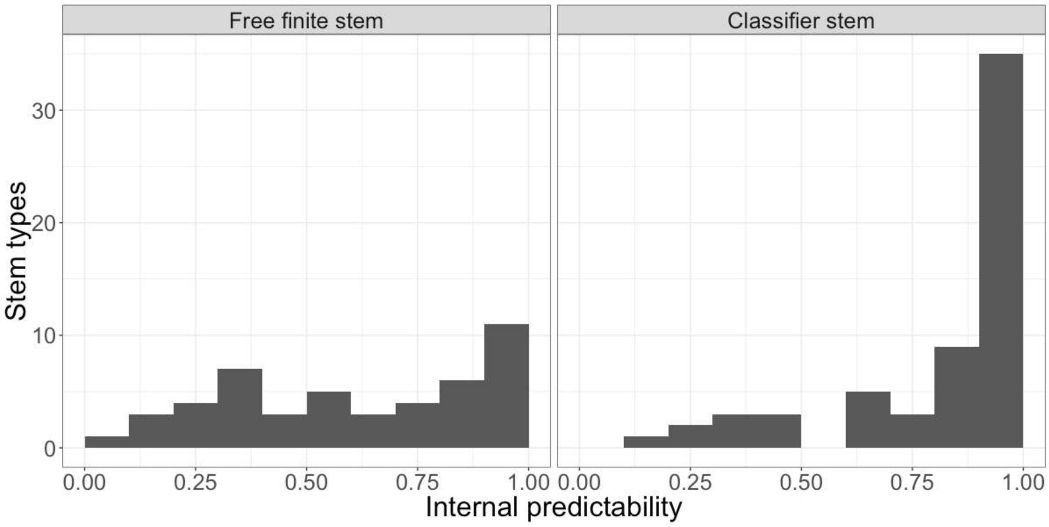

Figure 9. Murrinhpatha complex verb internal predictability, in free finite stem (recursive-word) and classifier stem (minimal-word) construction types.

Figure 10 shows overall entropy and mutual information values for the two construction subtypes. Recursive-word verbs have lower internal predictability IP(Ffs|C) = 1.64/2.45 = 0.67 compared to minimal-word verbs IP(Cls|C) = 2.38/2.82 = 0.84. This again supports the hypothesised link between selectional restrictiveness and internal predictability. The effect found in Murrinhpatha goes in the same direction as that in German, and the magnitude of the difference between construction types is very similar in the two languages.

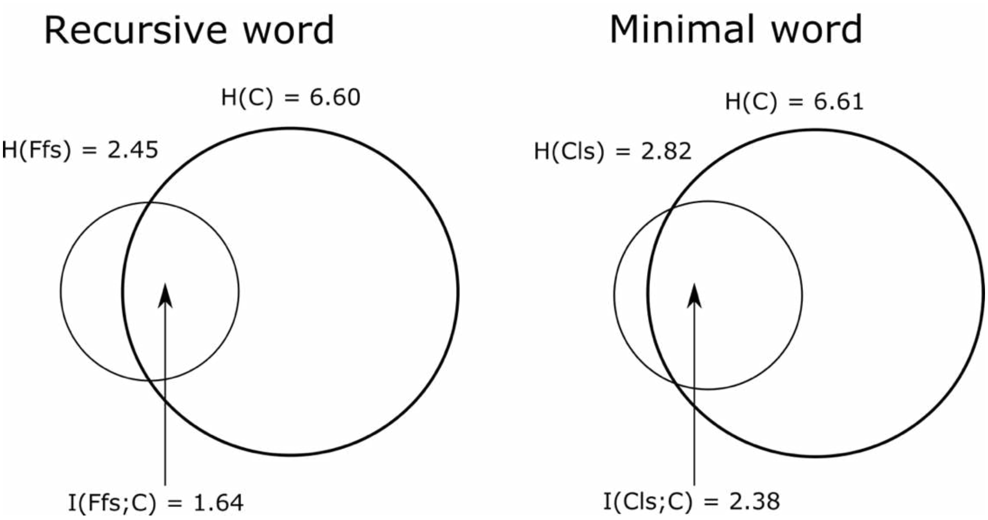

Figure 10. Independent entropies and mutual information for Murrinhpatha complex verb types: recursive-word (free finite stem) and minimal-word (classifier stem).

8. Summary and discussion

In this study I have illustrated an information-theoretic approach to wordhood, using a measure of internal predictability. This is defined as the proportion of one variable's independent entropy that is accounted for by mutual information with another variable. I illustrated the distribution of predictability among individual lexical elements, as well as the weighted average measure across all attested elements in the construction type. My hypothesis was that standard criteria for distinguishing complex words from phrases will correlate with internal predictability, and in this study I tested the method on selectional restrictiveness in complex verbs. The results here support the hypothesised relationship, though it remains to be seen whether the relationship will also stand for other construction types, and for other wordhood criteria.

In German complex verbs, where lexical verb stems are mostly free elements, there is a phrasal subtype combining the stem with a particle, and a recursive-word subtype combining it with a prefix. Internal predictability is substantially higher in the recursive-word type, in line with my hypothesis. Furthermore, a minimal-word type, combining a prefix with a selectionally restricted stem, has higher internal predictability again, suggesting a cumulative effect where both elements are selectionally restrictive. The German case study also shows that absolute mutual information is not higher in the word construction type, highlighting the importance of normalising mutual information against the independent entropy of the grammaticalised element.

In Murrinhpatha complex verbs, where lexical coverbs are mostly restricted morphemes, one subtype adds a promiscuous finite stem and is thus a recursive word, while the other subtype adds a selectionally restricted classifier and is thus a minimal word. Internal predictability is substantially higher in the minimal word type, again illustrating an association of selectional restrictiveness with internal predictability.

Figure 11 illustrates the internal predictability gradient for German and Murrinhpatha complex verb construction types. Recall that zero represents complete statistical independence of parts in a construction, and one represents complete dependence of one part on the other (§4.3). The German minimal word is the most internally predictable construction type, followed by the Murrinhpatha minimal word. Recursive word types, found in both languages but with different free elements [M-W] vs [W-M], are somewhat less internally predictable. The German phrase type is the least internally predictable of all. In this study I do not attempt to compare construction types between languages, as this introduces too many other dimensions of variance.

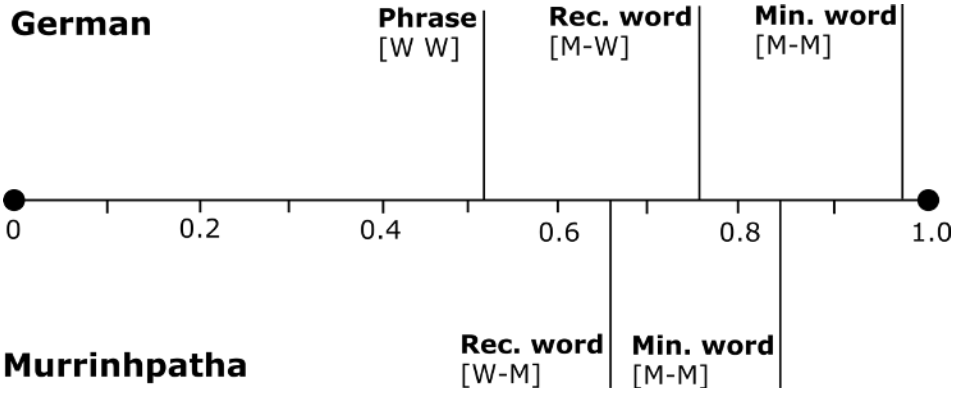

Figure 11. Complex verb subtypes on the internal predictability scale.

8.1. Selectional restrictiveness and cognitive representation

Why should a construction type with a selectionally restricted element be more internally predictable than a similar type with a promiscuous element? One possible explanation can be found in the idea of linguistic cognition as probabilistic 'chunking' (Christiansen & Chater 2015), given certain assumptions about the relationship between chunking and construction types.

The basic idea of chunking is that elements that predictably occur together tend to be parsed and produced as a holistic unit, while elements with more unpredictable relationships are parsed discretely, and produced by online composition. The interplay of these two options, often labelled 'storage' versus 'productivity', is central to a number of linguistic traditions including Cognitive Grammar (Blumenthal-Dramé 2012; Langacker 2017; Divjak 2019: 131ff.), and research on type and token frequencies (e.g. Baayen 1993; Hay 2002; Yang 2005; Bybee 2006; Barðdal 2008; Plag & Baayen 2009; Boyd & Goldberg 2011). O'Donnell (2015) proposes a sophisticated probabilistic model of the competition between storage and productivity, which accounts for a number of empirical facts about English past tense and noun derivations. To simplify the model somewhat, elements that occur in many rare combinations have stronger discrete representation, and elements that occur in a few highly frequent combinations are more strongly represented as part of those combinations. This has a clear relationship to conditional entropy, which is associated with large sets of rare combinations.

Where selectional restrictiveness might come into this picture is in the intrinsically higher conditional entropy of promiscuous elements, when measured beyond a single construction type. If an element is selectionally restricted to one construction type, then its conditional entropy in that construction type is its total conditional entropy in the language as a whole. But a promiscuous element has additional sources of conditional entropy, in other construction types. For a theory associating entropy with discrete representation, the question is to what extent this parsing bias is 'localised' to entropy within a particular construction type, or involves a more 'global' bias across all construction types. Does the representation of *vor-lassen* depend just upon the entropies of *vor* and *lassen* in the particle verb construction, or is it also influenced by their entropies in prepositional phrases, verb phrases etc? The former possibility would bias discrete processing according to local entropy, while the latter would bias discrete processing according to global entropy.

A global basis for discrete processing is favoured by one of the results in this study: the relevance of simple verb frequency to predictability in German particle verbs (§6.2). We saw that although stems in the particle verb construction generally have high particle entropy, low-frequency stems showed a different pattern with much more predictable particle selection. Since infrequent elements tend to have lower entropy, this supports the hypothesis that an element's entropy outside a construction type is in some way linked to its entropy within the construction type. I am not aware of any empirical or modelling work that pursues this hypothesis, but it would be a natural extension the current study, and O'Donnell's (2015) model, both of which treat one construction type at a time.

8.2. Consequences for the problem of wordhood

Given its limited scope, this study does not bring us closure on the problem of wordhood. What is achieved here is the articulation of a mathematical formula, internal predictability, which I hypothesise to be associated with several facets of the word concept. To illustrate, I have demonstrated its association with selectional restrictiveness in German and Murrinhpatha complex verbs. Despite the success of this experiment, I do not propose internal predictability as a 'measure of wordhood'. More modestly, I expect that wordhood properties will turn out to be correlated with internal predictability in many cases, though there may also be exceptions. My aspirations for the method are tempered by Wray's (2015) suggestion that the word concept is inherently vague:

> Orthographic practices encourage and perpetuate the expectation of clear and replicable boundaries between words. But perhaps they only fool us into believing that writing depicts what we already know, when in fact they are defining and marshalling aspects of a less tangible knowledge. Suppose the notion of word were inherently vague—too vague for us to feel comfortable about. How would theory handle that? (Wray 2015: 733)

If the word is an inherently vague concept, then the linguist's 'problem of wordhood' derives from the false hope that something intangible can be formulated in a rigorous technical fashion (Wittgenstein 1953: 19e). As proposed in other studies (Bickel et al. 2009; Haspelmath 2011: 64; Bickel & Zúñiga 2017; Tallman 2020a; Tallman 2020b), this suggests that progress in wordhood research will involve fuzzy or probabilistic associations between word-like properties, circumventing the impasse of criterial non-alignment. The information-theoretic formulation proposed in this study takes a slightly different approach, using a mathematical measure that is logically independent of standard wordhood criteria, but which I hypothesise may relate to several of them as a general tendency. It is hoped that this method will prove fruitful in future research, where probabilistic associations are seen as a worthy goal, and theoretical linguistics is not saddled by folk linguistics.

References

Aikhenvald, Alexandra. 2006. Serial verbs constructions in a typological perspective. In Aikhenvald, A. Y. & Dixon, R. M. W. (eds.), *Serial Verb Constructions: a cross-linguistic typology*, 1–68. Oxford, U.K.: Oxford University Press.

Attneave, Fred. 1959. *Applications of information theory to psychology: A summary of basic concepts, methods and results*. New York: Holt, Rinehart & Winston.

Baayen, Harald. 1993. On frequency, transparency and productivity. In Booij, Geert & van Marle, Jaap (eds.), *Yearbook of Morphology 1992* (Yearbook of Morphology), 181–208. Dordrecht: Springer Netherlands.

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3). 436–461.

Bannard, Colin & Matthews, Danielle. 2008. Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychological Science* 19(3). 241–248.

Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic*. Amsterdam: Benjamins.

Bauer, Laurie. 2017. *Compounds and compounding*. Cambridge: Cambridge University Press.

Belica, Cyril & Kupietz, Marc & Lüngen, Harald & Perkuhn, Rainer & Schächtele, Anna. 2014. *DeReWo – Corpus-based lemma and word form lists*. Leibniz Institute for the German Language. (https://www1.ids-mannheim.de/s/corpus-linguistics/projects/methods-of-analysis/corpus-based-lemma-and-word-form-lists.html?L=1) (Accessed April 30, 2020.)

Bickel, Balthasar & Banjade, Goma & Gaenszle, Martin & Lieven, Elena & Paudyal, Netra Prasad & Rai, Ichichha Purna & Rai, Manoj & Rai, Novel & Stoll, Sabine. 2007. Free prefix ordering in Chintang. *Language* 83(1). 43–73.

Bickel, Balthasar & Hildebrandt, Kristine A. & Schiering, Rene. 2009. The distribution of phonological word domains: a probabilistic typology. In Grijzenhout, Janet (ed.), *Phonological Domains: Universals and Deviations*, 47–78. Berlin: Mouton de Gruyter.

Bickel, Balthasar & Zúñiga, Fernando. 2017. The "word" in polysynthetic languages: Phonological and syntactic challenges. In Fortescue, Michael & Mithun, Marianne & Evans, Nicholas (eds.), *The Oxford handbook of polysynthesis*, 158–185. Oxford: Oxford University Press.

Biskup, Petr & Putnam, Michael & Smith, Laura Catharine. 2011. German particle and prefix verbs at the syntax phonology interface. *Leuvense Bijdragen* 97. 106–135.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*. Berlin, Boston: De Gruyter Mouton.

Blythe, Joe. 2009. *Doing referring in Murriny Patha conversation*. University of Sydney. (PhD thesis.)

Booij, Geert & Van Kemenade, Ans. 2003. Preverbs: an introduction. In Booij, Geert & Van Marle, Jaap (eds.), *Yearbook of Morphology 2003* (Yearbook of Morphology), 1–11. Dordrecht: Springer Netherlands.

Boyd, Jeremy K. & Goldberg, Adele. 2011. Learning what not to say: the role of statistical preemption and categorization in "a"-adjective production. *Language* 81(1). 1–29.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning* 34(1). 71–105.

Bresnan, Joan & Mchombo, Sam A. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13(2). 181–254.

Brinton, Laurel J. & Traugott, Elizabeth Closs. 2005. *Lexicalization and language change*. Cambridge: Cambridge University Press.

Bruening, Benjamin. 2018. The lexicalist hypothesis: Both wrong and superfluous. *Language* 94(1). 1–42.

Bybee, Joan L. 2006. From usage to grammar: the mind's response to repetition. *Language* 82. 711–733.

Christiansen, Morten H. & Chater, Nick. 2015. The now-or-never bottleneck: A fundamental constraint on language. *The Behavioral and Brain Sciences* 39. 1–52.

Coupé, Christophe & Oh, Yoon Mi & Dediu, Dan & Pellegrino, François. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* 5(9). eaaw2594.

Cover, Thomas A. & Thomas, Joy A. 2002. *Elements of information theory*. Second edition. London: Wiley.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Culbertson, Jennifer & Schouwstra, Marieke & Kirby, Simon. 2020. From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language* 96(3).

Di Sciullo, Anna-Maria & Williams, Edwin. 1987. *On the definition of word*. Cambridge, MA: MIT Press.

Divjak, Dagmar. 2019. *Frequency in language: Memory, attention and learning*. Cambridge: Cambridge University Press.

Dixon, R.M.W. & Aikhenvald, Alexandra Y. 2002. Word: a typological framework. In Dixon, R. M. W. & Aikhenvald, Alexandra (eds.), *Word: A cross-linguistic typology*, 1–41. Cambridge: Cambridge University Press.

Dodd, Bill & Eckhard-Black, Christine & Klapper, John & Whittle, Ruth. 2003. *Modern German grammar: A practical guide*. Second edition. London: Routledge.

Eisenberg, Peter. 2013. *Grundriss der deutschen Grammatik Band 1: Das Wort*. Stuttgart: J.B. Metzler.

Ellis, Nick C. & Ferreira–Junior, Fernando. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3). 370–385.

Futrell, Richard & Qian, Peng & Gibson, Edward & Fedorenko, Evelina & Blank, Idan. 2019. Syntactic dependencies correspond to word pairs with high mutual information. *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 3–13. Paris, France: Association for Computational Linguistics.

Geertzen, Jeroen & Blevins, James P. & Milin, Petar. 2016. Informativeness of linguistic unit boundaries. *Italian Journal of Linguistics* 28(1). 25–48.

Gibson, Edward & Futrell, Richard & Piantadosi, Steven T. & Dautriche, Isabelle & Mahowald, Kyle & Bergen, Leon & Levy, Roger. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23(5). 389–407.

Gijn, Rik van & Zúñiga, Fernando. 2014. Word and the Americanist perspective. *Morphology* 24(3). 135–160.

Goddard, Cliff. 1985. *A grammar of Yankunytjatjara*. Alice Springs: Institute for Aboriginal Development.

Gotelli, Nicholas J. & Chao, Anne. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. *Encyclopedia of biodiversity* 195–211.

Gries, Stefan Th. 2013. 50-something years of work on collocations. *International Journal of Corpus Linguistics* 18(1). 137–166.

Hafer, Margaret A. & Weiss, Stephen F. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10(11). 371–385.

Harris, Zellig S. 1955. From phoneme to morpheme. *Language* 31(2). 190–222.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.

Haspelmath, Martin. 2015. Defining vs diagnosing linguistic categories: a case study of clitic phenomena. In Blaszczak, Joanna & Klimek-Jankowska, Dorota & Migdalski, Krysztof (eds.), *How categorical are categories*. Berlin: Mouton de Gruyter.

Haspelmath, Martin. 2020. The morph as a minimal linguistic form. *Morphology* 30(2). 117–134.

Hay, Jennifer. 2002. From speech perception to morphology: Affix ordering revisited. *Language* 78(3). 527–555.

Hillert, Dieter & Ackerman, Farrell. 2002. Accessing and parsing phrasal predicates. In Dehé, Nicole & Jackendoff, Ray & McIntyre, Andrew & Urban, Silke (eds.), *Verb-particle explorations*. Berlin, Boston: De Gruyter Mouton.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276.

Langacker, Ronald W. 2017. Entrenchment in Cognitive Grammar. In Schmid, Hans-Jörg (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (Language and the Human Lifespan Series), 39–56. Boston, MA, US: De Gruyter Mouton.

Lester, Nicholas A. & Moscoso del Prado Martín, Fermín. in prep. Does language need syntax? Quantitative evidence from an extreme case.

Los, Bettelou & Blom, Corrien & Booij, Geert & Elenbaas, Marion & Kemenade, Ans van. 2012. *Morphosyntactic change: A comparative study of particles and prefixes*. Cambridge: Cambridge University Press.

Mansfield, John Basil. 2015. Morphotactic variation, prosodic domains and the changing structure of the Murrinhpatha verb. *Asia-Pacific Language Variation* 1(2). 162–188.

Mansfield, John Basil. 2016. Intersecting formatives and inflectional predictability: How do speakers and learners predict the correct form of Murrinhpatha verbs? *Word Structure* 9(2). 183–214.

Mansfield, John Basil. 2019. *Murrinhpatha morphology and phonology*. Berlin: De Gruyter Mouton.

Mansfield, John Basil & Blythe, Joe & Nordlinger, Rachel & Street, Chester. 2020. *Murrinhpatha morpho-corpus*. (langwidj.org/Murrinhpatha-morpho-corpus)

Matthews, Danielle & Bannard, Colin. 2010. Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive Science* 34(3). 465–488.

McDonald, S. A. & Shillcock, R. C. 2001. Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech* 44(Pt 3). 295–323.

McGregor, William. 2002. *Verb classification in Australian languages*. Berlin: Mouton de Gruyter.

Menzel, Wolfgang. 2019. *The Hamburg Dependency Treebank*. (http://hdl.handle.net/11022/0000-0000-7FC7-2) (Accessed March 2020.)

Mithun, Marianne. 2020. Where is morphological complexity? In Gardani, Francesco & Arkadiev, Peter M. (eds.), *Morphological complexity*. Oxford: Oxford University Press.

Montemurro, Marcelo A. & Zanette, Damián H. 2011. Universal entropy of word ordering across linguistic families. *PLoS ONE* 6(5).

Mugdan, Joachim. 1994. Morphological units. In Asher, R.E. (ed.), *The encyclopedia of language and linguistics*, 2543–2553. Oxford: Pergamon Press.

Müller, Stefan. 2002. Syntax or morphology: German particle verbs revisited. In Jackendoff, Ray & McIntyre, Andrew & Urban, Silke (eds.), *Verb-particle explorations*, 119–140. Berlin: De Gruyter.

Nordlinger, Rachel. 2015. Inflection in Murrinh-Patha. In Baerman, Matthew (ed.), *The Oxford handbook of inflection*, 491–519. Oxford: Oxford University Press.

Nordlinger, Rachel. 2017. The languages of the Daly River region (Northern Australia). In Fortescue, Michael & Mithun, Marianne & Evans, Nicholas (eds.), *Oxford handbook of polysynthesis*. Oxford: Oxford University Press.

O'Donnell, Timonthy J. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. Cambridge, MA: MIT Press.

Packard, Jerome L. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge: Cambridge University Press.

Pellegrino, François & Coupé, Christophe & Marsico, Egidio. 2011. A cross-language perspective on speech information rate. *Language* 87(3). 539–558.

Plag, Ingo & Baayen, Harald. 2009. Suffix ordering and morphological processing. *Language* 85(1). 109–152.

Ramscar, Michael & Port, Robert F. 2016. How spoken languages work in the absence of an inventory of discrete units. *Language Sciences* (Action, Culture, and Metaphor in Language Use) 53. 58–74.

Rice, Sally & Libben, Gary & Derwing, Bruce. 2002. Morphological representation in an endangered, polysynthetic language. *Brain and Language* 81(1–3). 473–486.

Russell, Kevin. 1999. The “word” in two polysynthetic languages. In Kleinhenz, Ursula & Alan, T. (eds.), *Studies on the phonological word*, 203–221. Amsterdam: John Benjamins.

Saenger, Paul. 1997. *Space between words: The origins of silent reading*. Stanford, CA: Stanford University Press.

Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace.

Schmid, Hans-Jörg & Küchenhoff, Helmut. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577.

Schultze-Berndt, Eva. 2003. Preverbs as an open word class in Northern Australian languages: synchronic and diachronic correlates. In Booij, Geert & van Marle, Jaap (eds.), *Yearbook of Morphology 2003*, 145–177. Amsterdam: Kluwer Academic Publishers.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423.

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30. 50–64.

Sosa, Anna Vogel & MacFarlane, James. 2002. Evidence for frequency-based constituents in the mental lexicon: collocations involving the word of. *Brain and Language* 83(2). 227–236.

Spencer, Andrew & Luis, Ana R. 2012. *Clitics: An introduction*. Cambridge: Cambridge University Press.

Street, Chester. 1987. *An introduction to the language and culture of the Murrinh-Patha*. Darwin: Summer Institute of Linguistics.

Street, Chester. 2012. *Murrinhpatha to English dictionary*. Wadeye Literacy Production Centre.

Tallman, Adam J. R. 2020a. Beyond grammatical and phonological words. *Language and Linguistics Compass* 14(2). e12364.

Tallman, Adam J. R. 2020b. Constituency and coincidence in Chácobo (Pano). *Studies in Language*.

Tallman, Adam J. & Wylie, Dennis & Adell, E. & Bermudez, N. & Camacho, G. & Epps, Patience & Woodbury, Anthony. 2018. Constituency and the morphology-syntax divide in the languages of the Americas: Towards a distributional typology. *21st Annual Workshop on American Indigenous Languages*. Santa Barbara: UCLA.

ten Hacken, Pius. 2017. Compounding in morphology. In Aronoff, Mark (ed.), *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press.

Tersis, Nicole. 2009. Lexical polysynthesis: Should we treat lexical bases aand their affixes as a continuum? In Mahieu, Marc-Antoine & Tersis, Nicole (eds.), *Variations on polysynthesis: The Eskaleut languages*, 51–64. Amsterdam: Benjamins.

van Marle, Jaap. 2002. Dutch separable compound verbs: Words rather than phrases? In Dehé, Nicole & Jackendoff, Ray & McIntyre, Andrew & Urban, Silke (eds.), *Verb-particle explorations*, 211–232. Berlin: De Gruyter.

Walsh, Michael. 1976. *The Murinypata language of north-west Australia*. Canberra: Australian National University. (PhD thesis.)

Widmer, Manuel & Auderset, Sandra & Nichols, Johanna & Widmer, Paul & Bickel, Balthasar. 2017. NP recursion over time: Evidence from Indo-European. *Language* 93(4). 799–826.

Williams, Edwin. 2007. Dumping lexicalism. In Ramchand, Gillian & Reiss, Charles (eds.), *The Oxford handbook of linguistic interfaces*, 353–381. Oxford: Oxford University Press.

Wittgenstein, Ludwig. 1953. *Philosophical investigations*. Third edition. Oxford: Blackwell. (Trans. Anscombe, G.E.M.)

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, Alison. 2015. Why are we so sure we know what a word is? In Taylor, John R. (ed.), *The Oxford handbook of the word*, 725–750. Oxford: Oxford University Press.

Yang, Charles. 2005. On productivity. *Linguistic Variation Yearbook* 5. 265–302.

Zwicky, Arnold M. & Pullum, Geoffry K. 1983. Cliticization vs. inflection: English N'T. *Language* 59(3). 502–513.

Appendix: Sample size effects on entropy and internal predictability

Entropy estimates can be highly inaccurate with small samples, and this is an important issue for any corpus study of lexical items, which by Zipf's law include many rare items. In this study the sample size effect is mitigated in two ways: firstly, but using the Chao-Shen entropy estimation method (Gotelli & Chao 2013), which corrects for small samples. Secondly, in the presentation of individual lexemes' predictability measurements (§6.1, §7.1), I exclude lexemes with less than 10 corpus tokens.

In this appendix I illustrate some effects of sample size on the estimation of complex verb entropy. I focus here on the German corpus data, which yielded a larger sample of 77,946 complex verb tokens. Comparing entropy estimates for smaller subsets of this data gives us some insight into the accuracy of the smaller Murrinhpatha sample, which consists of 6041 complex verb tokens.

Firstly, I show the effect of different sample sizes on estimating the preverb entropy of individual verb stems. This is done by drawing repeated independent samples from the full dataset.[20] Figure 1 shows particle entropy estimates for lexical stems appearing in the phrase construction, using the three stems illustrated in Table 7 of the main paper: *arbeiten*, *streichen*, *werfen*. Entropy estimates are on the y-axis, and sample size is on the x-axis (which is on a square-root scale). Cho-Shen entropy estimates are shown as heavier dots, and empirical entropy estimates as lighter crosses. For both methods, estimates have a high degree of variance with smaller samples, and gradually converge as the sample size increases. The Chao-Shen method both over- and under-estimates entropy in small samples, but importantly, estimates tend to cluster around the central value converged upon in larger samples. Empirical entropy estimates, on the other hand, systematically under-estimate entropy at smaller sample sizes.

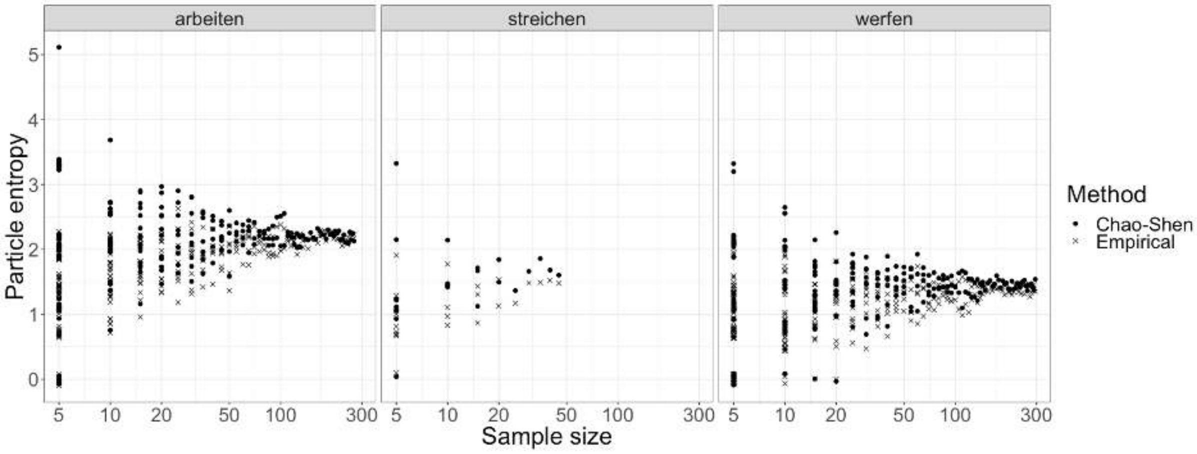

Figure 1. Particle entropy estimates for three German verb stems, using different corpus sample sizes.

In my presentation of preverb predictability among individual verb stems (§6.1, §7.1), a minimal token threshold of 10 was selected to mitigate estimation inaccuracy, while also

[20] See script sample-ent.R

including as many verb stems as possible. As shown in Figure 1, Chao-Shen estimates with only 10 tokens can be somewhat inaccurate, though estimates cluster towards the true value. The three stems shown here are among those with higher token counts (between 50–300), but as is typical with Zipfian lexical distributions, many stems have far fewer tokens. Therefore the preverb entropy estimates shown for individual verb stems in the main paper will have variable accuracy, according to token count, with N≥10 set as a floor to avoid the most egregious errors.

Figure 2 shows prefix entropy estimates for three verb stems in the word-type construction. All have very low prefix entropy. At smaller sample sizes the figure shows some massive over-estimates, which occur when a small sample happens to include one of the rare prefix combinations. However the vast majority of small-sample estimates are in fact zero, i.e. quite accurate. Over-plotting of points obscures the predominance of accurate estimates, but regression lines (dashed for Chao-Shen, solid for empirical) have been added to show the overall accuracy. The stem *reichen* only ever occurs with the prefix *er-* in our sample, and therefore all estimates are zero.

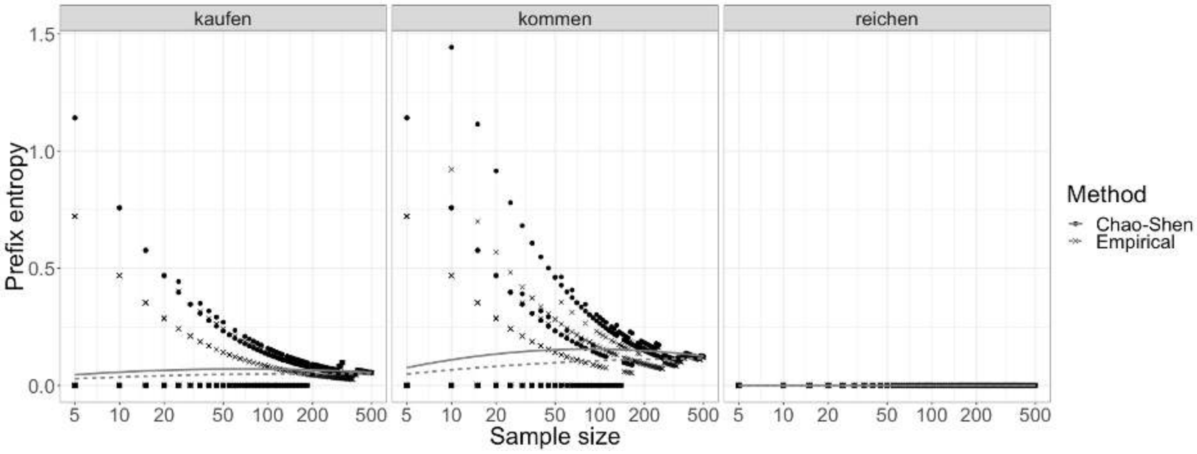

Figure 2. Prefix entropy estimates for three German verb stems, using different corpus sample sizes.

In the overall measures of construction type internal predictabililty (IP) (Figure 5 in the main paper), all verb stems are included irrespective of token frequency. This gives a more complete picture of predictability in the construction type, since rare lexemes are an intrinsic part of corpus distributions. Importantly, IP is a weighted average across verb stems, and therefore takes into account token frequency (i.e. verb stem probability), in a way that is not evident in the individual lexeme figures. Highly frequent stems, with more accurate entropy estimates, have a greater influence on IP. Low-frequency stems, with less reliable entropy estimates, each have a very small influence on IP.

Finally, it is worth considering the effect of the total sample size on IP, especially since Murrinhpatha provided a much smaller sample. Figure 3 shows IP measures for different sized independent samples of the German complex verb dataset. Again both Chao-Shen and empirical estimates are shown. Cho-Shen estimates (dots) converge to a stable value by around 10,000 complex verb tokens. Empirical estimates (crosses) overestimate IP, especially in the more unpredictable phrase construction. Given that 6041 tokens were available for Murrinhpatha complex verbs, and assuming that the laws of sample size would apply similarly to Murrinhpatha as to German, we can see that Chao-Shen estimates for Murrinhpatha are likely to be accurate within a few percentage points.

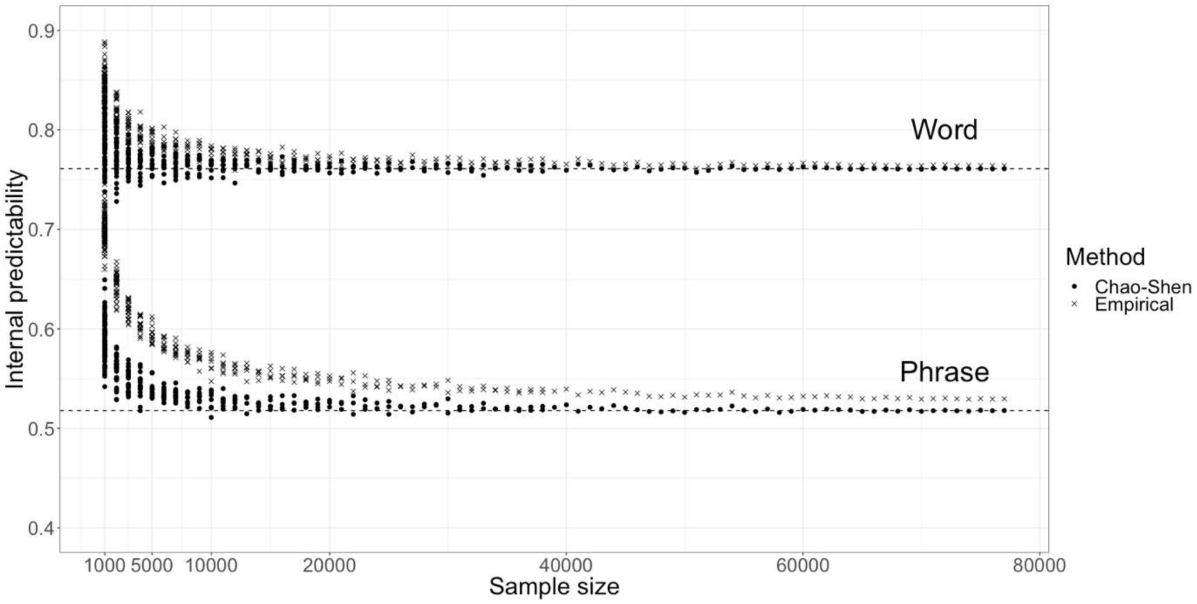

Figure 3. Internal predictability estimates for German complex verb construction types, using different corpus sample sizes.