Measuring Morphological Fusion Using Partial Information Decomposition

Michaela Socolof
McGill University/Mila

Jacob Louis Hoover
McGill University/Mila

Richard Futrell
University of California, Irvine

Alessandro Sordoni
Microsoft Research

Timothy J. O'Donnell
McGill University/Mila
Canada CIFAR AI Chair

Abstract

Morphological systems across languages vary when it comes to the relation between form and meaning. In some languages, a single meaning feature corresponds to a single morpheme, whereas in other languages, multiple meaning features are bundled together into one morpheme. The two types of languages have been called agglutinative and fusional, respectively, but this distinction does not capture the graded nature of the phenomenon. We develop a mathematically precise way of characterizing morphological systems using partial information decomposition, a framework for decomposing mutual information into three components: unique, redundant, and synergistic information. We show that highly fusional languages are characterized by high levels of synergy.

1 Introduction

Languages are, to a large extent, systematic; there are predictable patterns in the way that meanings are mapped to forms. However, languages differ when it comes to the nature of the relation between meaning and form. This variability is particularly apparent in the domain of morphology, and underlies the distinction between so-called **agglutinative** and **fusional** languages (von Humboldt, 1825; Greenberg, 1960). The two types of languages differ in the extent to which multiple **units of meaning** are expressed by a single morpheme. In this paper, a unit of meaning simply refers to a semantic (or grammatical) feature such as *plural* or *accusative*. Highly agglutinative languages have words that are built up of clearly separable morphemes, each of which corresponds to an individual unit of meaning. The relationship between meaning and form in these languages is thus highly systematic. On the other hand, highly fusional languages fuse together multiple units of meaning into a single affix that cannot be decomposed in any obvious way, and so are less systematic.

Hungarian		Russian	
Meaning	*Form*	*Meaning*	*Form*
cat-SG-DAT	macská-∅-nak	cat-SG-DAT	кот-у
cat-PL-DAT	macská-k-nak	cat-PL-DAT	кот-ам
cat-SG-TERM	macská-∅-ig	cat-SG-GEN	кот-а
cat-PL-TERM	macská-k-ig	cat-PL-GEN	кот-ов

Table 1: In Hungarian (left), every unit of meaning tends to correspond to a morpheme hence the meaning-form relationship is systematic. On the contrary, in Russian (right) such correspondence cannot be found. We aim to quantify the degree of systematicity in meaning-form relations across morphological systems.

In Table 1, we illustrate the meaning-form mapping for words in Hungarian (an agglutinative language) and Russian (a fusional language). In the Hungarian paradigm, *singular*, *plural*, *dative*, and *terminative* each always correspond to a single morpheme, which is the same across contexts. In Russian, the affixes package together multiple units of meaning and cannot be decomposed: there are no morphemes that individually correspond to *singular*, *plural*, *dative*, or *genitive*—rather, the form of the suffix depends on multiple meaning units.

The agglutinative versus fusional distinction captures a core intuition about the different ways meaning can correspond to morphological form, but the distinction is binary and therefore does not characterize the graded nature of the phenomenon (Greenberg, 1960)—that is, the fact that different languages (and, indeed, specific domains within a language) show varying degrees of fusion. In this paper, we take an information-theoretic approach to quantifying systematicity in meaning-form relations across morphological systems.

The core insight we draw upon is that meanings can contribute information about a linguistic form in three different ways. First, a unit of meaning can provide information about the form that no other unit of meaning provides. This is called **unique** information. Second, a unit of meaning can provide the exact same information about the form that another unit of meaning provides. This is called **redundant** information. Third, a unit of meaning can, in combination with some other unit of meaning, jointly provide information that is not provided by either on its own. This is called **synergistic** information. Going on these definitions, we expect fusional languages to have a higher relative amount of synergy than agglutinative languages.

We argue that these three kinds of information in morphological systems correspond precisely to existing notions of unique, redundant, and synergistic information in the information theory literature. In particular, the **Partial Information Decomposition** (PID) framework, introduced by Williams and Beer (2010), decomposes the mutual information between a target variable and two (or more) source variables into unique, redundant, and synergistic information. This decomposition of mutual information into three components makes up the **information profile** of a system. When we take form to be the target variable and the individual meaning features to be the source variables, the information profile gives the amount of information conveyed individually, concurrently, or jointly, by units of the meaning about form. Crucially, two systems can have equal mutual information between meaning and form, but different information profiles, corresponding to different degrees of morphological fusion. Therefore, PID offers a mathematically precise way of placing morphological systems along an agglutinative-to-fusional spectrum.

In summary, our contributions are as follows. We use the PID framework to develop a novel measure of the systematicity of meaning-form mappings in morphological systems. To validate our method, we first carry out two simulations using artificial languages for which we can control the degree of morphological fusion. We show that languages possessing a low relative amount of synergistic information are the most systematic. Finally, we apply the decomposition to morphological systems in 22 real languages, successfully recapitulating existing linguistic categorizations in a graded way.

2 Partial information decomposition

2.1 The problem

A fundamental property of language is that linguistic forms depend on the meaning being communicated. Information theory gives us a way of quantifying this dependence, with **mutual information**, a measure of how much one random variable informs us about another random variable (Shannon, 1948; Fano, 1961). Let M and F be discrete random variables representing meaning and form, respectively. The mutual information $I(M; F)$ between M and F can be expressed as:

$$I(M;F) = \sum_{m\in M}\sum_{f\in F} P(m,f)\log\frac{P(m,f)}{P(m)P(f)}. \tag{1}$$

M	$\rightarrow$	F
aa		00
ab		01
ba		10
bb		11

M	$\rightarrow$	F
aa		00
ab		01
ba		11
bb		10

Table 2: (left) An example of a fully *systematic*, or one-to-one, code, in which each variable in F is informative about a variable of M. (right) This code is *less systematic* because the value of each F variable depends on more than one M variable. Here $F = \text{CNOT}(M)$. Both codes have $I(M;F) = 2$ bits.

In a linguistic system, both the meaning and the form have internal structure, and it is the relationship between subparts of these structures that we are interested in. We therefore define both M and F as ensemble random variables, made up of sets of random variables corresponding to the individual units of meaning and form. As an example, consider the two toy languages in Table 2. In both languages, M is an ensemble random variable made up of two binary random variables (one for each column). Similarly, F is composed of two binary random variables. Assuming a uniform distribution on the inputs, the mutual information between M and F in both languages is 2 bits, since it takes 2 bits of information on average to communicate about the meaning. However, the mutual information on its own does not tell us whether the relation between meaning and form variables is one-to-one, many-to-one, etc. In the language on the left, one variable (i.e., column) on the meaning side fully determines each variable on the form side. In the second language, both meaning variables are needed to correctly predict each form variable.

Since mutual information does not tell us how the information is distributed among the pieces of meaning and form, we want to decompose mutual

Collection	Associated information about T
$\{S_1\}$	Unique (U_1) of S_1
$\{S_2\}$	Unique (U_2) of S_2
$\{S_1\}\{S_2\}$	Redundancy ($R_{1,2}$) of S_1 and S_2
$\{S_1, S_2\}$	Synergy ($S_{1,2}$) of S_1 and S_2

Table 3: Collections and associated information quantities for the case of two source variables about a target variable T.

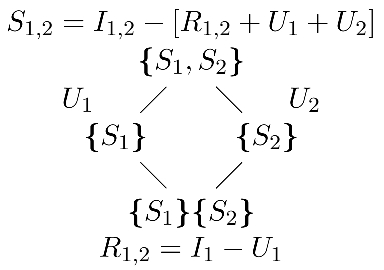

Figure 1: Partial information lattice for the case of two source variables. The equations at each node are abbreviated versions of equations (2)–(4), showing how to solve for redundant, unique, and synergistic information, starting at the bottom of the tree.

information based on how each meaning variable contributes information—on its own, redundantly with other variables, or jointly with other variables.

2.2 Partial Information Decomposition

Decomposing mutual information requires extending traditional information theory to handle multivariate interactions, such as that between two or more meaning variables that jointly provide information about a form variable. Williams and Beer (2010)'s PID framework provides an influential solution to the decomposition problem; we briefly summarize the framework here.

Williams and Beer (2010) set up the problem as a decomposition of the ways that *source* variables provide information about a *target* variable. Consider the simple case of two source variables S_1 and S_2 and a target variable T. Let a *collection* be a grouping of one or more nonzero subsets of source variables such that none of the subsets is a superset of any other. There are four such collections: $\{S_1\}$, $\{S_2\}$, $\{S_1\}\{S_2\}$, and $\{S_1, S_2\}$. Each collection is then associated with a particular quantity of information, summarized in Table 3.[1] The sum of these quantities is the mutual information $I(S_1, S_2; T)$. For the sake of brevity, we will use U, R, and S as shorthand for unique, redundant, and synergistic information, and subscripts to indicate information about T from source variables S_1 and/or S_2.

Williams and Beer show that the collections can be naturally structured into a partially-ordered lattice, shown for the case of two source variables in Figure 1. At the bottom is the information provided redundantly by S_1 and S_2. The next level up is the information provided uniquely by S_1 and the information provided uniquely by S_2. At the top is the information jointly contributed by S_1 and S_2, i.e., the synergy. An important feature of the lattice is that the mutual information between a set of sources and the target is the sum of all nodes below and including the collection consisting of that particular set of sources. This means that the values at all nodes in the entire lattice add up to the mutual information provided by the two sources S_1 and S_2 about the target, as expressed in Equation 2. It also means that the mutual information between a single source S_1 and the target is made up of the unique information in S_1 plus whatever information is redundant between S_1 and S_2, expressed in Equation 3 (and the same for S_2, in Equation 4).

$$I(S_1, S_2; T) = R_{1,2} + U_1 + U_2 + S_{1,2} \tag{2}$$

$$I(S_1; T) = R_{1,2} + U_1 \tag{3}$$

$$I(S_2; T) = R_{1,2} + U_2 \tag{4}$$

These equations all have a mutual information term on the left, which we have a definition for and can therefore compute. However, we do not at this point know how to compute any of the terms on the right, so we have a system of three equations with four unknowns, which we cannot solve.

Gutknecht et al. (2020), building on Williams and Beer (2010), show that with a definition of either redundant information or unique information, it is possible to solve the system of equations 2–4 for the remaining variables using a Möbius inversion function to move recursively up the lattice. Much work in the PID literature has focused on formulating an independent definition for redundant or unique information (e.g., Williams and Beer, 2010; Bertschinger et al., 2014; Finn and Lizier, 2018; Makkeh et al., 2021). A number of solutions have been proposed, and there is as yet not total consensus on the "best" measure. Below, we will adopt one such measure, which is both common in the literature and intuitive for our application—that of Bertschinger et al. (2014).

[1] This can be generalized to an arbitrary number of source variables. See Williams and Beer (2010) for details.

Bertschinger et al. give an independent definition for unique information. Their measure is based on the intuition that the unique information of S_1 should reflect the information about T which is *only* available from S_1, regardless of the choice of S_2. This is operationalized by adversarially computing the minimum possible conditional mutual information $I_Q(S_1 : T \mid S_2)$, minimizing over all possible joint distributions $Q(S_1, S_2, T)$ that have the same marginals as the true distribution P:

$$U_1 = \min_{Q \in \Delta_P} I_Q(S_1; T \mid S_2) \tag{5}$$

where

$$\begin{aligned}\Delta_P = \{Q \in \mathfrak{P}(S_1, S_2, T) \mid \\ &\textstyle\sum_{s_2' \in \mathcal{S}_2} Q(s_1, s_2', t) = P(s_1, t) \wedge \\ &\textstyle\sum_{s_1' \in \mathcal{S}_1} Q(s_1', s_2, t) = P(s_2, t) \\ &\forall t \in \mathcal{T}, s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2\}\end{aligned}$$

where $\mathfrak{P}$ is the set of all joint distributions.

The Bertschinger et al. (2014) formulation of PID is known to give intuitive results on a number of canonical example distributions; for example in the mapping from the second variable of meaning M to the second variable of form F in the codes of Table 2, we get a unique information of 1 for the fully systematic example and 0 for the less systematic example. In Section 3.2 we define a measure of morphological fusion based on the Bertschinger et al. (2014) formulation.

3 Methods

We compute PID between meaning and form of noun paradigms in suffixing languages from UniMorph (Sylak-Glassman, 2016), which contains annotated morphological data for 167 languages using a universal schema. An example paradigm is in Table 4. All of the languages in our experiment have noun paradigms with exactly two non-stem meaning feature categories: CASE and NUMBER.

3.1 Defining meaning and form variables

In order to compute the partial information decomposition, we first need to define our source and target random variables. Since we are interested in how each component of meaning contributes individually or jointly to determining linguistic forms, we treat meaning variables as our sources and form variables as our targets.

Meaning	*Form*
cat-NOM-SG	кот
cat-NOM-PL	коты
cat-GEN-SG	кота
cat-GEN-PL	котов
cat-DAT-SG	коту
cat-DAT-PL	котам
cat-INS-SG	котом
cat-INS-PL	котам
cat-ESS-SG	коте
cat-ESS-PL	котах

Table 4: Subset of paradigm for the Russian noun кот.

Source Meaning Variables Consider the morphological paradigm for the Russian noun кот in Table 4, which consists of inflected forms of the word paired with their grammatical information. For our source variables, we treat each meaning feature category (CASE, NUMBER, and the stem) as a random variable with values that range over the possible feature values (e.g., *nominative* or *singular* for CASE and NUMBER, respectively).

Target Form Variables In order to define our target random variables, it is necessary to decompose the suffixes in some way, since treating the entire suffix as a target would not allow us to investigate its degree of internal agglutination or fusion. To define random variables over forms, we adopt an alignment-based approach, breaking up the suffixes into morphological slots and treating each slot as a random variable whose values range over the different aligned sequences that appear in the slot. We perform the alignment using LingPy's morphological aligner (List and Forkel, 2021). In order to compute PID, it is necessary for the number of random variables to be consistent across all words in the paradigm, so we pad empty slots with a dummy character. The number of random variables, then, is determined by the word with the longest suffix in the paradigm. In the majority of alignments, each slot ends up containing a one- or two-character sequence. An example alignment of several Russian words and the resulting form slots is shown in Table 5.

Our application differs from the original PID formulation in that we are dealing with multiple target variables. In Section 3.2 we propose an expectation-based approximation of PID for the joint distribution over multiple targets. In what follows, meaning random variables are denoted by $\mathcal{M} = \{M_1, \ldots, M_n\}$, while the form random variables are denoted by $\mathcal{F} = \{F_1, \ldots, F_m\}$. M_1 and

M_1	M_2	M_3	F_1	F_2	F_3	F_4
cat	GEN	SG	КОТ	а	-	-
cat	DAT	PL	КОТ	а	м	-
cat	INS	PL	КОТ	а	м	и

Table 5: Random variable structure for three word forms in Russian.

F_1 represent the stem's meaning (e.g., *cat*) and the stem's form (e.g., КОТ), respectively (Table 5).

3.2 Computing PID

Within each language, we compute PID on each noun's paradigm individually. Our motivation for treating each noun separately is that in many languages, morphological paradigms vary based on features of particular stems. For example, in a language with a gender distinction, the combination of meaning features *accusative*+*plural* might be expressed differently on masculine versus feminine nouns. We argue that this is not relevant to the notion of agglutinative versus fusional that we are interested in. If *accusative* and *plural* are expressed by separate morphemes in masculine as well as feminine nouns, then the fact that their specific forms vary with gender does not make the language any less agglutinative. It would be possible to extend our approach to handle stem-specific features in a dataset that made this information available, but since UniMorph does not annotate these features, we proceed with computing PID on each noun separately. With this approach, we are essentially treating the stem as a proxy for any stem-specific information, and conditioning all of our probability distributions, and thus our PID quantities, on the stem. In Appendix A, we give an example of how aligning multiple UniMorph-style paradigms without accounting for stem-specific features can obscure systematic regularities.

Since we treat each paradigm separately, the form and meaning variables M_1 and F_1 corresponding to the stem are generally uninteresting to us, as they remain constant throughout each paradigm. This approach is equivalent to computing the information-theoretic quantities in PID conditioned on the stem variables M_1, F_1. We are left with exactly two source variables, which correspond to CASE and NUMBER, and $m-1$ target variables. In what follows, for simplicity, we relabel the meaning variable corresponding to CASE as M_1 and the one corresponding to NUMBER as M_2.

We are now challenged with computing the PID between M_1, M_2 and the set of form variables $F \in \mathcal{F} - \{F_1\}$. We do this by taking the expectation of the PID quantities of these two variables with each form variable as target, separately. We first compute the PID between the two meaning variables M_1, M_2 and one of the form variables $F \in \mathcal{F} - \{F_1\}$ and obtain the values of unique $U_{1,2\to F}$, redundant $R_{1,2\to F}$, and synergistic $S_{1,2\to F}$ information, where we have made the dependence on the particular form variable F explicit in the subscript for clarity. We then normalize these quantities by the mutual information $I(M_1, M_2; F)$ to obtain the relative amount of each type of information. For each combination of meaning variables M_1, M_2 and one form variable F, the proportion of unique, redundant, and synergistic information that M_1 and M_2 give about F is:

$$\bar{U}_{1,2\to F} := \frac{U_{1\to F} + U_{2\to F}}{I(M_1, M_2; F)} \tag{6}$$

$$\bar{R}_{1,2\to F} := \frac{R_{1,2\to F}}{I(M_1, M_2; F)} \tag{7}$$

$$\bar{S}_{1,2\to F} := \frac{S_{1,2\to F}}{I(M_1, M_2; F)} \tag{8}$$

To compute the total average amount of each type information for a given language, we average these quantities across the full set of form variables of the paradigm and across paradigms. Let $h \in \{\bar{U}, \bar{R}, \bar{S}\}$. The average amount of information of type h in the language is:

$$\bar{h} = \frac{1}{|\mathcal{N}|} \sum_{n\in\mathcal{N}} \frac{1}{|\mathcal{F}|} \sum_{F\in\mathcal{F}-\{F_1\}} h_{1,2\to F} \tag{9}$$

where $\mathcal{N}$ is the set of paradigms in our dataset. We give pseudo-code for this process in Appendix B.

We use an implementation of Bertschinger et al. (2014)'s measure given in Wollstadt et al. (2018). Computing PID for the full set of nouns in every language is computationally intensive, so instead we repeatedly subsample the paradigms for $|\mathcal{N}| = 10$ different nouns, randomly selected, from each language. We do this 100 times per language.

4 Experiments

We first validate that Bertschinger et al. (2014)'s PID measure captures the phenomenon we are interested in by running the measure on noun paradigms in two sets of artificial languages. After validating the measure, we then apply the PID framework to noun paradigms in 22 real languages, showing that the proportion of synergy characterizes the degree of fusion in a linguistic system.

4.1 Artificial languages — intuition

Our artificial languages are generated based on the intuition that in a highly agglutinative language, each inflection corresponds to a single unit of meaning, whereas in a highly fusional language, each inflection corresponds to a combination of meanings. We operationalize these intuitions by generating random languages where inflections are sampled either conditioned on single meaning features (agglutinative) or sampled conditioned on pairs of meaning features (fusional). We test on a set of very simple artificial languages as well as a set of artificial languages that were generated to match a number of statistical properties of real languages in our dataset, and thus control for a variety of linguistic phenomena.

4.2 Artificial languages — simple

We generated fifteen very simple artificial languages. In each language, the noun paradigms had six cases and three numbers. The first five languages were "agglutinative," where the suffixes were two-character strings, with one character independently generated conditionally on one meaning variable. A second set of five "fusional" languages were generated such that each suffix was a random two-character string sampled conditionally on both meaning features. Finally, as a sanity check we generated a set of five baseline languages that were intended to be as synergistic as possible. Under the Bertschinger et al. (2014) measure, XOR is a maximally synergistic boolean function. Therefore, the control languages were generated using XOR. Each suffix was a single character long with two possible realizations corresponding to the boolean values output by the XOR function and given by $F(\text{case}, \text{number}) = (\text{case} \in C)$ XOR $(\text{number} \in N)$, where C and N are random nonempty proper subsets of the possible case and number values, respectively. The PID results for these artificial languages in Figure 2 confirm that the measure captures the differences between these artificial languages as expected. All five agglutinative languages have 100% unique information, while the XOR languages have majority synergistic information. The fusional languages fall in the middle, with a proportion of synergy between 20% and 40%.

4.3 Artificial languages — linguistic controls

In our second experiment, we validate our measures using more linguistically-realistic artificial languages that are matched to real languages for specific properties, such as the size of the character vocabulary, phonotactic restrictions, and average suffix length, as well as other properties that may correlate with the agglutinative/fusional distinction. We do this by generating agglutinative and fusional versions of existing languages.

We began by selecting six languages whose noun paradigms are given in UniMorph. Each of the languages in UniMorph is labelled as either agglutinative or fusional, based on information from linguistic analysis; two of our chosen languages (Hungarian and Turkmen) are labeled as agglutinative, and the remaining four (Ukrainian, German, Latin, and Northern Sami) are labeled as fusional. For each language, we trained a 3-gram model on all the language's inflected nouns, to approximate the language's phonotactics, and used this model to generate artificial paradigms for that language. For each language we sampled fifty artificial agglutinative paradigms and fifty artificial fusional paradigms following the sampling scheme outlined above. To sample an artificial fusional paradigm, we used the 3-gram model to generate random suffixes for the stem, jointly conditioned on case and number. To generate an artificial agglutinative paradigm, we generated independent strings for each value of case and number, and concatenated them (in either order, but consistent within a paradigm). For both types, we sampled suffixes with a range of lengths to roughly match that of the suffixes in the real language. The PID results are shown in Figure 3. These results confirm that our PID measure captures the difference between agglutinative and fusional paradigms in the expected way: The agglutinative versions of the languages had proportionally less synergistic and more unique information than the fusional versions, regardless of which type of language they were generated from.

4.4 Real languages

We investigate whether PID provides a way of measuring morphological fusion by computing PID on noun paradigms from 22 languages in UniMorph. Seven of our languages are labeled as agglutinative, and the remaining ones as fusional. Our results are given in Figure 4, which shows the relative amount of unique, redundant, and synergistic information for each language. Languages with an asterisk and solid black outline are those that were labelled as

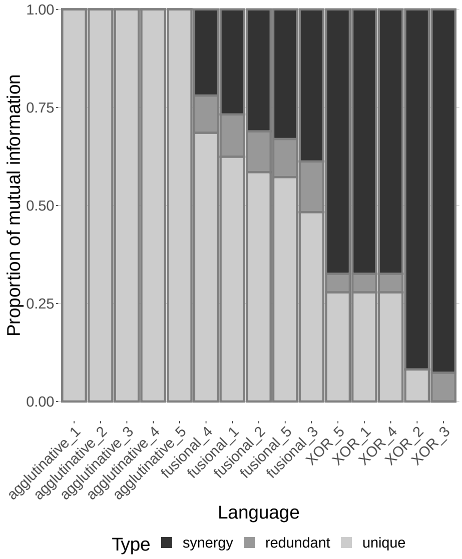

Figure 2: Results of partial information decomposition on noun paradigms in baseline artificial languages. The languages are sorted by relative amount of synergy.

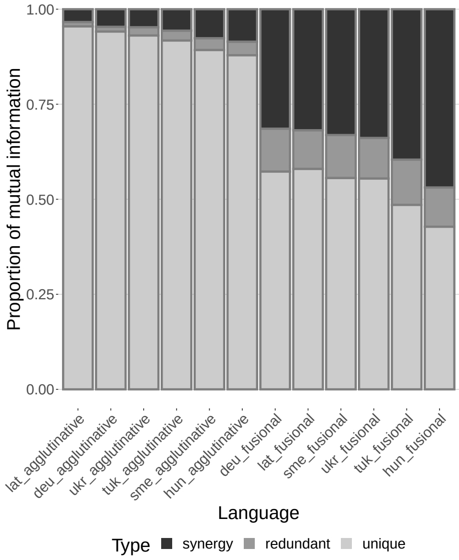

Figure 3: Results of partial information decomposition on noun paradigms in linguistically-controlled artificial languages. The languages are sorted by relative amount of synergy.

agglutinative in UniMorph. We find that the seven agglutinative languages fall on the side of lowest synergy, though there were also a few fusional languages that had low synergy.

As baselines for the PID measure, we present (1) a plot of the average amount of mutual information between meaning and form in the individual nominal paradigms across the 22 languages in our experiment (Figure 5 in Appendix C), and (2) a plot of the average number of suffix slots in each language (Figure 6 in Appendix C). The baselines suggest that high mutual information and high suffix length are often present in agglutinative languages, but our artificial experiments reveal that when we control for these factors, PID successfully captures the amount of fusion present in a system.

5 Discussion

Our results suggest that PID does indeed capture the spectrum between agglutinative and fusional. We also find that there is more unique information overall than redundant or synergistic information, which points to an overall high level of systematicity in morphology. Redundant information makes up the smallest proportion of information overall, suggesting that morphological systems are not particularly redundant. This raises the question of whether other domains in language show similar levels of redundancy, and how the low amount of redundancy should be accounted for.

While PID seems to be able to capture morphological fusion, it is important to note that the agglutinative/fusional classification system is very coarse—when we apply a single label to each language, we miss fine-grained distinctions such as the fact that different domains within a language can have different degrees of fusion. For this reason, we believe that when evaluating any measure of fusion, it is best to examine the actual paradigms. Let us consider a paradigm from Latin, shown in Table 6. Latin falls far to the right side of the spectrum, and we can see in the paradigm that there is a lack of systematicity among the suffixes. For example, the *-s* in column F_5 appears with both singular and plural, and across four different cases. The PID values for each combination of variables are given in Table 7. We can see that for every combination of two meaning variables and one form variable, there is more synergy than any other type of information.

6 Related work

There is a growing literature on information-theoretic approaches to problems in morphology and syntax. One line of work looks at the trade-off between the surprisal of a linguistic form and the

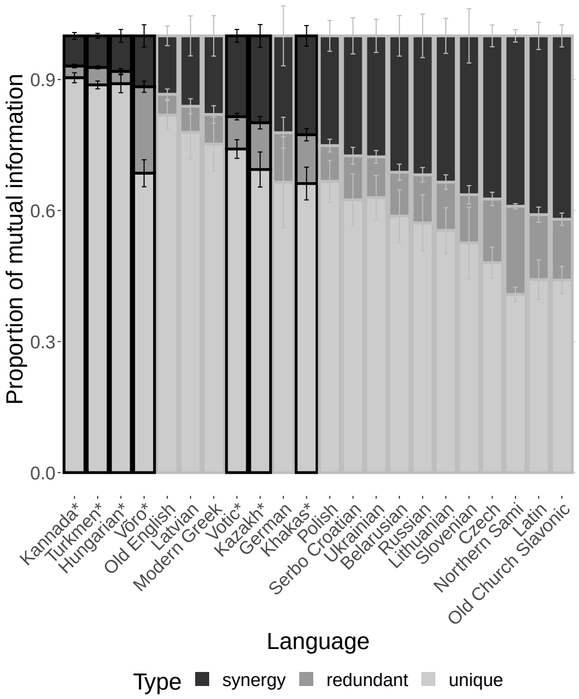

Figure 4: Results of partial information decomposition on noun paradigms in 24 languages. The languages are sorted by relative amount of synergy. Asterisks and dark borders represent languages labeled as agglutinative in UniMorph.

M_1	M_2	M_3	F_1	F_2	F_3	F_4	F_5
nur	NOM	SG	nur	-	-	u	s
nur	NOM	PL	nur	ū	s	-	-
nur	GEN	SG	nur	ū	s	-	-
nur	GEN	PL	nur	-	-	uu	m
nur	DAT	SG	nur	-	-	u	ī
nur	DAT	PL	nur	i	b	u	s
nur	ACC	SG	nur	-	-	u	m
nur	ACC	PL	nur	ū	s	-	-
nur	ABL	SG	nur	ū	-	-	-
nur	ABL	PL	nur	i	b	u	s
nur	VOC	SG	nur	-	-	u	s
nur	VOC	PL	nur	ū	s	-	-

Table 6: Random variable structure for a Latin noun.

s_1	s_2	t	U_1	U_2	R	S
M_2	M_3	F_2	0.323	0.135	0.16	0.865
M_2	M_3	F_3	0.445	0.39	0.014	0.61
M_2	M_3	F_4	0.355	0	0.136	0.833
M_2	M_3	F_5	0.689	0	0.095	1

Table 7: PID values (unique, redundant, synergistic) for the Latin paradigm in Table 6, unnormalized.

time it takes to produce (Pimentel et al., 2021); the trade-off between surprisal and memory in accounting for word and morpheme order cross-linguistically (Hahn et al., 2021); and mutual information as a measure of the relationship between grammatical gender and co-occurring words (Williams et al., 2021). Accounting for patterns of word and morpheme order across languages using information theory has yielded a variety of proposed measures (Hahn et al., 2020; Dyer et al., 2021).

Closely related to our work is Rathi et al. (2021), which proposes a measure of *informational fusion* in morphology, based on Wu et al. (2019)'s definition of morphological irregularity. Let ℓ be a lexeme, σ a semantic feature combination, and w a surface form. *Informational fusion* is defined as:

$$\phi(w) = -\log p(w \mid \mathcal{L}_{-\sigma}, \sigma, \ell) \qquad (10)$$

Informational fusion is a measure of the surprisal of a surface form given the rest of the paradigm. Unlike the PID approach, which involves segmenting the suffix and finding the information profile of each subpart, informational fusion is computed with respect to un-segmented forms, and does not make reference to individual morphemes. PID gives us a way of investigating the exact question we are interested in—to what extent do units of the meaning individually or jointly contribute information about individual units of the form? We use PID to get at the fine-grained distinctions between information profiles, an approach that we believe can be extended to study compositionality more generally.

7 Conclusion

We have proposed a novel way of characterizing morphological systems cross-linguistically, using partial information decomposition. PID allows us to decompose the mutual information between meaning and form into three distinct components: unique, redundant, and synergistic information. We argued that the relative amount of synergistic information provides a mathematically precise and intuitive measure of the degree of fusion in a morphological system. We carried out a study on noun paradigms, demonstrating the promise of this approach in this specific domain. Our study applies PID at the level of morphemes, and suggests extensions to word- and sentence-level domains, potentially leading to a more general theory of compositionality. We see PID as an exciting tool for investigating the information profile of any system in which meaning features are expressed by linguistic forms.

Acknowledgments

We thank the Montreal Computational & Quantitative Linguistics Lab and the Language Processing Group at the University of California, Irvine for helpful feedback. We also gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, the Fonds de Recherche du Québec, and the Canada CIFAR AI Chairs Program.

References

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. 2014. Quantifying unique information. *Entropy*, 16:2161–2183.

William Dyer, Richard Futrell, Zoey Liu, and Greg Scontras. 2021. Predicting cross-linguistic adjective order with information gain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 957–967, Online. Association for Computational Linguistics.

Robert M. Fano. 1961. *Transmission of information: a statistical theory of communications*. MIT Press.

Conor Finn and Joseph T. Lizier. 2018. Pointwise partial information decomposition using the specificity and ambiguity lattices.

Joseph H. Greenberg. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26(3):178–194.

Aaron J. Gutknecht, Michael Wibral, and Abdullah Makkeh. 2020. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *CoRR*, abs/2008.09535.

Michael Hahn, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 9117(5):2347–2353.

Wilhelm von Humboldt. 1825. Über das Entstehen der grammatischen Formen und ihren Einfluss auf die Ideenentwicklung. In *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin: Aus den Jaren 1822 und 1823*, pages 401–430.

Johann-Mattis List and Robert Forkel. 2021. A Python library for historical linguistics. version 2.6.9.

Abdullah Makkeh, Aaron J. Gutknecht, and Michael Wibral. 2021. Introducing a differentiable measure of pointwise shared information.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Neil Rathi, Michael Hahn, and Richard Futrell. 2021. An information-theoretic characterization of morphological fusion. In *EMNLP*.

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

John Sylak-Glassman. 2016. The composition and use of the Universal Morphological Feature Schema (UniMorph Schema).

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damian Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159.

Paul L. Williams and Randall D. Beer. 2010. Nonnegative decomposition of multivariate information.

Patricia Wollstadt, Joseph T. Lizier, Raul Vicente, Conor Finn, Mario Martinez-Zarzuela, Pedro Mediano, Leonardo Novelli, and Michael Wibral. 2018. Idtxl: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software*, 4(34):1081.

Shijie Wu, Ryan Cotterell, and Timothy J. O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Association for Computational Linguistics*.

A

As an illustration of how computing PID on the full set of noun paradigms without accounting for stem-conditioned features can obscure the patterns, consider the following toy paradigms:

M_1	M_2	F_1
NOM	SG	a
NOM	PL	b
ACC	SG	a
ACC	PL	b

Table 8: Toy language, noun 1.

M_1	M_2	F_1
NOM	SG	a
NOM	PL	a
ACC	SG	c
ACC	PL	c

Table 9: Toy language, noun 2.

In the first paradigm, M_2 uniquely determines F_1. In the second paradigm, M_1 uniquely determines F_2. For both nouns, there is 1 bit of unique information, and no redundant or synergistic information. Thus all of the mutual information between meaning and form in this language is unique. However, if we compute PID on the full set of forms without conditioning on noun 1 and noun 2, we get 0.66 bits of unique information, 0.016 bits of redundant information, and 0.077 bits of synergistic information. This irregularity comes from the fact that the suffix *-a* serves different functions for the different nouns, but the PID measure considers both types of *-a* to be the same realization of F_1. Crucially, this means we can get synergy in a language whose individual paradigms do not actually have any synergy.

B

```
def compute_pid(paradigm):
  N = paradigm.num_nouns
  F = paradigm.num_F # num form variables
  M = 2              # num meaning variables (2)
  V = numpy.zeros((N, M + F))
  # fill the matrix of values
  vtoi = dict()
  for n in range(N):
    for m in range(M):
      # convert string value of s to int
      value = paradigm[n].meaning[m]
      if value not in vtoi:
        vtoi[value] = len(vtoi)
      V[n, m] = vtoi[value]
    for f in range(F):
      # convert string value of f to int
      value = paradigm[n].form[f]
      if value not in vtoi:
        vtoi[value] = len(vtoi)
      V[n, f] = vtoi[value]
  # compute PID for each target var
  bar_u, bar_r, bar_s = 0, 0, 0
  for f in range(F):
    u, r, s, mi = pid(
        V, sources=[0, 1], target=2 + f
    )  # Bertschinger's PID using IDTXL
    bar_u += u / mi # avg. unique
    bar_r += r / mi # avg. redundant
    bar_s += s / mi # avg. synergy
  return bar_s/F, bar_u/F, bar_r/F
```

Listing 1: Python-style pseudo-code for computing relative PID quantities for a given paradigm.

C

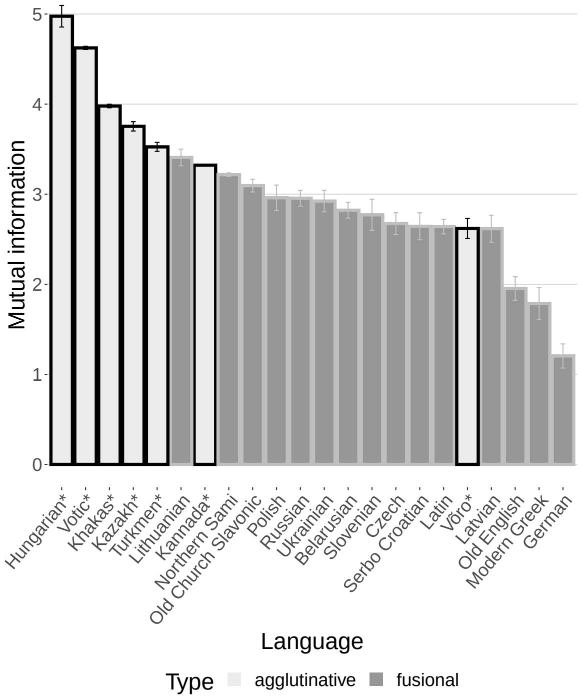

Figure 5: Average amount of mutual information between meaning and form in the nominal paradigms of 22 languages. Asterisks and dark borders represent languages labeled as agglutinative in UniMorph.

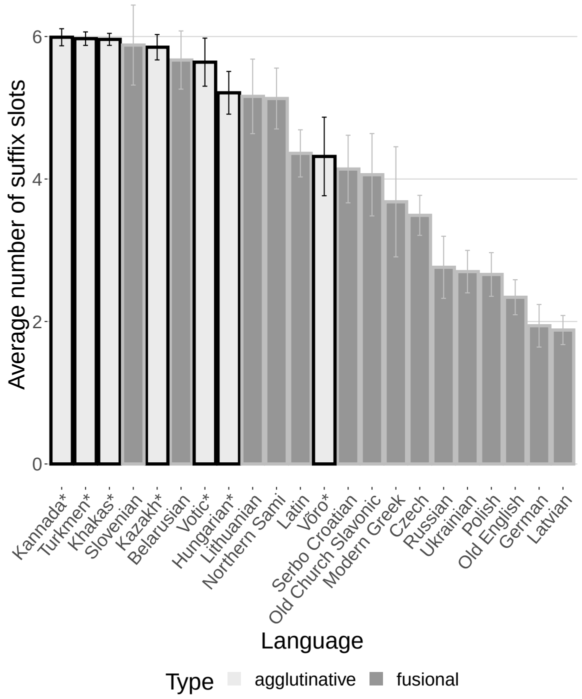

Figure 6: Average suffix length in the nominal paradigms of 22 languages. Asterisks and dark borders represent languages labeled as agglutinative in UniMorph.