Partial compositionality

Michaela Socolof

Department of Linguistics

McGill University, Montreal

June 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Contents

Abstract

This thesis investigates the relationship between compositionality and idiosyncrasy in language. Many types of linguistic expressions have been described as exceptions to compositionality, with idioms being the most well-documented examples. Idioms—expressions like *spill the beans*—can be characterized by two important properties. First, they exemplify non-conventional word meaning; for example, the words *spill* and *beans* in this idiom seem to carry particular meanings—something like *reveal* and *information*, respectively—which are different from the conventional meanings of these words in other contexts. Second, there is a contingency relationship between words in an idiom. It is the specific combination of the words *spill* and *beans* that has come to carry the idiomatic meaning. I conduct a corpus study showing that idioms (as well as non-idiomatic phrases) vary in the extent to which they display these two properties. I then zoom in on conventionality as a proxy for the compositional contribution of a word to a larger phrase. A sentence production study investigates whether processing time is sensitive to compositionality and finds that the effect is negligible.

The central theoretical contribution of this thesis is a new way of formalizing compositionality using an information-theoretic framework called Partial Information Decomposition (PID). This framework allows one to measure the precise amount of compositional information that a linguistic unit contributes to a larger expression. A PID-based definition of the *degree of compositionality* of an utterance is proposed, and the measure is evaluated on a corpus on idioms. The framework is then extended to capture variation in morphological systematicity.

Résumé

Cette thèse jette un regard sur la relation entre la compositionnalité et l'idiosyncrasie du langage. Il y a plusieurs types d'expressions linguistiques qui ont été décrit comme étant non-compositionnel, incluant les expressions idiomatiques. Les expressions idiomatiques tels que « spill the beans » ont deux propriétés importantes. La première est que la définition du mot dans l'expression idiomatique est différente de leurs définitions d'origine. Cette propriété est étroitement reliée à la compositionnalité. Par exemple, les mots « spill » et « beans » dans cette expression idiomatique veulent dirent « révéler » et « secret » respectivement. La deuxième est une relation de contingence entre les mots contenu dans l'expression. Autant le mot « spill » que « beans » doit être présent pour que l'expression ait un sens idiomatique.

Dans cette thèse, je conduis quatre expériences. La première expérience démontre que les expressions idiomatiques varient selon les deux propriétés décrit plut tôt. Ma deuxième expérience examine la vitesse à laquelle les gens prononcent des phrases idiomatiques et non idiomatiques. Les résultats démontrent que la compositionnalité n'a pas un grand effet sur la vitesse d'élocution. Je propose une nouvelle façon de formuler la compositionnalité en utilisant la théorie de l'information. Spécifiquement, j'utilise un cadre mathématique appelé « *partial information decomposition* », qui nous laisse mesurer la compositionnalité de toute expression. J'évalue la nouvelle mesure de compositionnalité en utilisant un corpus contenant des expressions idiomatiques. Finalement, j'utilise le « *partial information decomposition* » pour mesurer la systématicité dans la morphologie de plusieurs langages.

Acknowledgments

The first person I'd like to thank is Tim O'Donnell, who is to a large extent responsible for how enjoyable and invigorating my PhD experience has been. From the two-hour conversation when I was deciding whether to do my PhD at McGill all the way to the end of the dissertation-writing process, Tim has been consistently supportive, encouraging, and inspiring. He has gone above and beyond in every way and is always generous with his time, energy, and insight. He's encouraged me to think big and be bold, modeling what it means to truly understand a topic, to delve deeper, to ask more questions. I'm in awe of his seemingly endless energy for rigorous discussion and grateful to him for encouraging me to follow the branches of my interests, even when they diverged into unrelated topics. It's been a joy to have an advisor who is both a great scientist and a great mentor, someone who helped guide me along the sometimes-murky path of research with enthusiasm and profound humanity. Tim challenged me to step outside my comfort zone, was receptive to my instincts (the ones that led to interesting places and the ones that didn't), and was always up for a discussion. His lab environment made it fun to do a PhD and to be stuck in the weeds of an idea. It's been a privilege and an absolute pleasure.

I'd like to thank the other members of my committee, Michael Wagner, Richard Futrell, and Alessandro Sordoni. My discussions with Michael always led to new insights, and his questions and suggestions made my thinking better. He was the one who suggested I look at idiom prosody as part of this project, and I'm so glad I did. I first got to know Richard through an information theory working group, and his knowledge and ideas have continued to add depth to my understanding of many of the topics in this dissertation. Alessandro has always been supportive and eager to learn about the linguistic phenomena I was interested in, while providing excellent ideas about the computational aspects. Our conversations frequently left me thinking about topics in new and

rewarding ways.

Many other people at McGill have contributed to this work and to my intellectual development as a whole. I'd like to thank Jackie Cheung, who co-supervised my first paper on idioms. Thanks to Jessica Coon for guiding me through the depths of trying to figure out the Georgian agreement system, even when things got hairy. Thanks to Bernhard Schwarz and Aron Hirsch for their support and collaboration on my first semantics project. Other faculty and postdocs in the department with whom I have had stimulating and fruitful discussions include Morgan Sonderegger, Meghan Clayards, Nico Baier, Eva Portelance, Lisa Travis, Luis Alonso-Ovalle, Junko Shimoyama, Siva Reddy, Martina Martinović, Charles Boberg, and Eva Portelance. A special thank you to Lisa Travis, who taught my first syntax class and made me fall in love with the subject, and Junko Shimoyama, who advised my undergraduate thesis. Thanks to Morgan Sonderegger and Michael Wagner for hiring me as an RA in the Montreal Language Modelling Lab all those years ago, and to Michael McAuliffe for helping me learn to code. My TAs during undergrad—Maayan Abenina-Adar, Brian Buccola, Dan Goodhue, Alanah McKillen—inspired and encouraged me, both directly and indirectly. Thanks to Andria de Luca and Giuliana Pannetta in the main office, who were always brilliant, helpful, and willing to indulge my harebrained schemes. Thank you to all the other Linguistics faculty and staff over the years for making the department such a great environment.

I've spent these past years surrounded by wonderful fellow students. Thank you to Emily Goodwin, friend and confidante. Our chats and bakes and visits over the years have meant so much to me. Thanks to Vanna Willerton for being an incredible host on my prospective student visit and always making the lab a blast. Thanks to lab members Ben LeBrun, Emi Baylor, Graham Adachi-Kriege, Greg Theos, and Yves Blain-Mostesano. A million thanks to my labmate and cohortmate, Jacob Hoover. I'm so lucky to have had you with me every step of the way. Outside the lab, I've greatly enjoyed my time and conversations with many other students in the department, especially Jonny Palucci, Justin Royer, and Will Johnston.

During my PhD I've had the opportunity to mentor some outstanding undergraduates. Béatrice Vallières helped with stimuli creation for the idiom production experiment in this thesis. Cypress Zufferli worked on a project extending the morphology study reported here. Thanks also to Chase Boles, with whom I spent many enjoyable hours thinking about conditional clauses in Igala.

Thank you to Kenny Smith for hosting me at the University of Edinburgh, and to everyone in

the Centre for Language Evolution and beyond for being so welcoming and helpful, especially Maisy Hallam, Elizabeth Pankratz, Kate McCurdy, and Zheng Zhao.

A detour, now, to the year I spent as a Baggett Fellow at the University of Maryland before starting my PhD. That year was instrumental in my development as a researcher and has informed much of my trajectory since. Thank you to Naomi Feldman for giving me the opportunity to do computational linguistics research when I was still new to the subject. Thank you to the grad students and postdocs who went out of their way to show me kindness and friendship during some of the toughest moments: Paulina Lyskawa, Aaron Doliana, Anouk Dieuleveut, Max Papillon, and especially Ted Levin.[1] Most importantly, I want to thank my fellow postbacs: Nancy Clarke, Jackie Nelligan, Hanna Muller, and Jon Burnsky. So many shenanigans, so much laughter. We were all reckoning with what we wanted our futures to look like, in academia or outside, which areas of linguistics we were interested in, whether we had it in us. The highs were high and the lows were low, and I'm grateful I got to spend that wild and whimsical year with you all.

Thank you to those outside academia who have made these years better. To Talia, Matt, Abbey, Yelim, Jess, Braedan, Quinn, Zoe, Rachael, Diego, Miklós, Alex, François, Maddie, Emma, Austin, Anne, Aase, Ivi, Nat, Amelie, Mathieu. To my parents, my sister, my grandparents. You all are the best.

To Montreal—for everything.

[1] And Cardboard Ted, of course.

Contribution of authors

I am the first author of all of the manuscripts in this thesis. Chapter 3 was published in the *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* and was co-authored by Jackie Chi Kit Cheung, Michael Wagner, and Timothy J. O'Donnell. I developed the research questions in collaboration with Timothy J. O'Donnell, and I conducted the experiments and analyzed the data. Jackie Chi Kit Cheung provided input on decisions regarding the corpus and implementation of the measures. Michael Wagner provided input on the theoretical discussion of idioms as well as help with the behavioral experiment software and design. I wrote the paper and received editorial feedback from all three co-authors.

Chapter 4 was co-authored by Timothy J. O'Donnell and Michael Wagner. Michael Wagner initially suggested looking at idiom prosody, and I developed the research questions and experiment in consultation with him and Timothy J. O'Donnell. I directed undergraduate student Béatrice Vallières in the creation of the stimuli, and I carried out the experiment and statistical analyses. For the statistical analyses and the interpretation of results, I consulted with both co-authors. I prepared the manuscript with editorial feedback from both.

Chapter 5 is work in preparation with Timothy J. O'Donnell. I was responsible for the original conception of the bounds analysis of the compositionality measure, and I collaborated with Timothy J. O'Donnell on formulating details of the implementation on corpus data. I carried out the experiments and data analysis, and I prepared the manuscript with his editorial feedback.

Chapter 6 was published in the *Proceedings of the 29th International Conference on Computational Linguistics* and was co-authored by Jacob Louis Hoover, Richard Futrell, Alessandro Sordoni, and Timothy J. O'Donnell. Through discussions on information theory with all four co-authors, I developed the idea of using Partial Information Decomposition to systematicity in morphology. I

designed and carried out the natural language experiment. Jacob Louis Hoover created the artificial languages for the baseline comparisons, and I performed the analyses on these languages. Richard Futrell contributed text to the section explaining Partial Information Decomposition, and Alessandro Sordoni wrote the pseudocode illustrating how the measure was computed. I was responsible for all other writing. All four co-authors provided editorial feedback.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The compositionality debate

The *Principle of Compositionality* has been at the heart of decades of debate in linguistics and related fields. The debate concerns whether natural language should be viewed as a compositional system, where compositionality captures the idea that we express complex meanings by combining smaller units with simpler meanings.

(1) *Principle of Compositionality*: The meaning of an expression is determined by the meanings of its parts and the way those parts are combined.

Those who argue that language is compositional typically describe a setup in which words behave like building blocks and are shuffled and recombined according to syntactic rules to create sentences with predictable meanings. Under this view, the principle of compositionality is usually understood to be augmented by the further property of *systematicity*, whereby words (or other linguistic units) contribute consistent meanings across all of the expressions in which they occur. Furthermore, proponents of this view argue that phrases appearing to violate compositionality, such as idioms, do not necessitate abandoning the central intuition. On the other side of the debate are those who believe that the pervasiveness of context dependence in language means that language should not be considered a compositional system.

In some sense, the divide between these positions is a false dichotomy—most researchers agree that language involves both systematicity and context dependence. In fact, language is commonly

described as being mostly, or somewhat, compositional. However, traditional frameworks for formalizing compositionality have obscured this common ground by characterizing compositionality as an all-or-nothing property, rather than as something that can partially hold. The most influential family of formalizations is built on Montague (1970)'s proposal that a compositional system is one where there is a *homomorphism* between expressions in a language and their meanings; under such a framework, compositionality is a property of a system as a whole (e.g., Montague, 1970; Janssen, 1986; Hendriks, 2001; Szabó, 2004). Yet compositionality need not be formalized as an all-or-nothing property; the central intuition could be made precise in a variety of ways, with the traditional framework offering just one possibility. For example, while the statement of compositionality above requires a deterministic relationship between the meaning of an expression and its parts/structure, there is no requirement that the relationship be systematic (though systematicity is frequently understood to be an important part of compositionality). In discussing compositionality, it is important to disentangle the two central questions that underlie the debate:

(2) **Central questions:**

a. What does (or should) the principle of compositionality refer to?

b. Is language compositional? (To what extent?)

Often, the second question takes precedence, overshadowing the fact that different assumptions have been made about the first. A common answer to the first question is that a language is compositional as long as some function determines the meaning of an expression based on the meanings of its parts and their syntax. This way of understanding compositionality is consistent with traditional symbolic approaches like those based on logics, lambda calculi, and formal grammars, and has been influential in linguistics in particular. Those who believe that language is compositional typically argue that the ability of individual words to be combined into larger phrases in predictable ways is a fundamental property of language, and that without it, language would not be learnable or productive (e.g., Fodor & Katz, 1964; Putnam, 1975; Montague, 1974; Partee, 1984; Pelletier, 2004). The learnability argument says that if the relationship between sentences and their meanings were completely arbitrary (as the relationships between morphemes/words and their meanings often are), then given the fact that we have so many millions of sentences to learn, the cognitive task would be overwhelming. The related notion of productivity refers to the fact that we can create sentences that

have never been said before, and the productivity argument says that when we encounter a sentence we have never heard before, we can infer its meaning based on our knowledge of how the words were used and combined in other sentences that we do know. For example, knowing the meanings of the words *spill* and *water*, and knowing how to combine a verb with a direct object, allows us to understand the phrase *spill water*. There are also predictable patterns among sentences whose meanings we understand, due to the fact that they are made from recombining parts from other sentences we understand. Thus, if we know how to obtain the meanings of "The dog wants chocolate" and "The cat wants milk," in virtue of the meanings of their parts and the syntax combining those parts, we should also be able to understand the sentence "The cat wants chocolate," since its words and its syntactic structure are familiar.

An example of this flavor of argument being made explicitly comes from Frege (1923), who points out that the ability to converge on the same (or a similar) meaning between speaker and listener would be inexplicable were it not for a computation that relies on systematically mapping parts of sentences to corresponding meanings. He writes,

> "It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a terrestrial being for the very first time can be put into a form of words which will be understood by someone to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thoughts corresponding to the parts of a sentence, so that the structure of the sentence serves as the image of the structure of the thought" (Frege, 1923, p. 53)

On the other hand, it is frequently argued that much, if not most, linguistic meaning depends on context, and that the existence of expressions with context-dependent meanings provides evidence against language being compositional. This view is particularly common among those who model language using continuous, rather than discrete, representations—e.g., neural networks. A common line of argument against compositionality is that nearly all instances of word usage have their meanings in some way constrained by context. Consider the sentence *Sally bounced the ball up and down.* What does *ball* mean in this sentence? It must be a type of ball that can be bounced—not, for example, a bowling ball—so there is a restriction on the meaning of ball that comes from the

context surrounding the verb *bounce*. This point has been made explicitly by Taylor (2008) and Langacker (1987), with Taylor (2008) using this fact to argue that the role of compositionality is often overstated as an axiom of language.[1] The argument about context dependence is captured well in the following quote from Lahav (1989):

> "What it is for a bird to count as red is not the same as what it is for other kinds of objects to count as red. For a bird to be red (in the normal case), it should have most of the surface of its body red, though not its beak, legs, eyes, and of course its inner organs. Furthermore, the red colour should be the bird's natural colour, since we normally regard a bird as 'really' red even if it is painted white all over. A kitchen table, on the other hand, is red even if it is only painted red, and even if its 'natural' colour underneath the paint is, say, white." (264).

Perhaps the most frequently-discussed instances of context dependence—and therefore purported counterexamples to compositionality—are idioms (e.g., J. Katz & Postal, 1963; Chomsky, 1980; van der Linden, 1992). In the sentence in (3), *spill the beans* means something close to "reveal the information" or "reveal the secret." Crucially, this meaning is not predictable solely from knowing the meanings of *spill*, *the*, *beans*, and the syntactic operation that combines a verb with a direct object.

(3) The doctor is threatening to spill the beans about the top-secret treatment.

Taylor (2008) makes a further argument against the idea that linguistic expressions are built compositionally, namely that a large portion of a speaker's knowledge of language consists of fixed or semi-fixed multiword expressions which, "by dint of their frequent use, have become entrenched in the minds of language users and which are not plausibly assembled afresh on each occasion of their use" (12). These would include idioms but also common phrases with more straightforward meanings, such as *back and forth* and *dual citizen*. Such phrases, he argues, behave similarly to individual words in that they are memorized with their corresponding meanings; even if they can be built compositionally, they are generally not computed this way in practice.

[1] Langacker (1987) in fact uses the term *partial compositionality* to describe the fact that much of linguistic meaning depends on context.

The debate about whether language is compositional is in some sense a debate about priorities. The central facts—that language involves systematicity and context dependence—are largely uncontroversial, but one or the other of these phenomenon might be more or less relevant depending on one's research questions. Those who believe that language is compositional argue that compositionality can be formalized in such a way that captures the apparent exceptions, and that abandoning compositionality would leave basic facts about language unaccounted for (such as that when we learn a new word, we can immediately use it in a variety of sentences in predictable ways). Furthermore, proponents of compositionality note that context dependence cannot be entirely unrestricted—that is, the size of the context that a meaning can depend on must be finite, given human memory limitations and the fact that language is learnable. On the other hand, those who believe language is not compositional argue that the size of the context that must be considered is larger than any linguistic unit such as word or constituent phrase, and in fact there is no single grain size of linguistic unit for which language can be said to be completely compositional. For instance, if we take words as the linguistic unit of interest, then idioms make language seem non-compositional, since the words in idioms do not contribute their canonical meanings. We could increase the size of our units to be phrases rather than words, but there would still be higher-level contextual information affecting the interpretation of those phrases. The same is true if we narrow our focus and take morphemes as the unit of interest. If we went down to an even smaller grain size—phonemes—language wouldn't look compositional at all, since a phoneme is generally not associated with the same meaning across contexts. The two sides of the debate can be understood as emphasizing different facts about language: that language is learnable, productive, and largely systematic, yet also exhibits large amounts of context dependence such that there is no single grain size of linguistic unit for which language is entirely compositional.

1.2 Idioms and partial meaning contribution

This thesis focuses primarily on idioms, using them as a test case to argue for a framework of partial (rather than all-or-nothing) compositionality. Idioms demonstrate an empirical phenomenon that has important consequences for discussions of compositionality, namely that words can share some aspect of their meaning across utterances while also being to some extent context dependent. In

the quote above from Lahav (1989), there is some degree of context dependence that distinguishes between the meanings of *red* in *red table* and in *red bird*, yet at the same time a significant portion of their meaning is shared (namely, the quality of the color being described). When it comes to idioms, this phenomenon of partial meaning contribution occurs in a subtler way. For most idioms, the idiomatic interpretation of a particular word shares some aspect of its meaning with the word's literal interpretation. For example, in the idiom *spill the beans*, the verb *spill* has a meaning similar to its literal meaning of releasing something that is not supposed to be released, but not entirely the same. There is also a part of the phrase's meaning that is idiosyncratic to the co-occurrence of *spill* and *beans*. Thus we can say that some of the word's meaning is invariant, and some is context dependent.

Related to this idea of partial meaning contribution is the work of Nunberg et al. (1994), which points out that some idioms are intuitively compositional in the sense that we assign metaphorical meanings to their parts and then put those parts together in a compositional way. The words in these idioms thus contribute part of their literal meaning—whatever part is shared with the metaphorical meaning—and the amount of literal contribution varies across idioms (e.g., *spill* contributes more of its literal meaning in *spill the beans* than *kick* does in *kick the bucket*—meaning "die"). Nunberg et al. (1994) further hypothesizes that the compositionality of an idiom determines which grammatical transformations it can undergo. In particular, the authors claim that the more compositional an idiom is, the more syntactically flexible it is, which is why one can say the idiom *spill the beans* in a passive construction ("The beans were spilled") but not the idiom *kick the bucket* (*"The bucket was kicked"). This contrast is striking because *spill the beans* and *kick the bucket* share the same basic structure—a verb and a direct object—yet we implicitly know that only one of the two can be passivized. This suggests that there is a difference in the way the two idioms are stored in the mind, allowing us to infer the correct generalizations about which syntactic configurations are possible. In order to test this and other hypotheses about the relationship between meaning contribution and grammatical structure, it is necessary to have a way of measuring the degree of compositionality of linguistic expressions.

Another area where we see partial meaning contribution from the literal words in an idiom is in aspectual meaning. McGinnis (2002) observes that idioms (1) preserve the aspectual properties associated with literal uses of their components parts, and (2) display the full range of aspectual

classes that are available for phrases of their syntactic profile. For example, the idiom *kick the bucket* means something close to "die," but it cannot appear in all of the aspectual configurations that the word *die* can; specifically, it shares the aspectual restrictions of transitive verbs with definite direct objects.

(4) Hermione was dying for weeks.

(5) *Hermione was kicking the bucket for weeks. (McGinnis, 2002)

The examples given in this section represent just a small fraction of the ways that partial meaning contribution presents in language. Handling the phenomenon in a formal framework has proven difficult, however, as traditional approaches do not offer an obvious path toward treating individual expressions as involving varying amounts of compositionality.[2]

The main aims of this thesis are twofold. First, the thesis presents an empirical investigation into the phenomenon of partial compositionality, looking at idioms as a test case. This investigation uses corpus data and behavioral experiments to characterize the role of partial compositionality in language as well as its effects on processing. Second, the thesis aims to capture partial compositionality formally, presenting a unifying framework (based in information theory) that allows us to measure the compositional and non-compositional contributions of any linguistic unit. A specific measure of the *degree of compositionality* of an expression is proposed; this measure allows us to place expressions along a spectrum of compositionality and isolate the contributions from each particular sub-unit. The proposed framework preserves the insights from both sides of the compositionality debate—that language is largely systematic, and that language involves a great deal of context dependence—while abstracting away from particular theories of meaning representation.

1.3 Information-theoretic concepts

This section introduces the basic concepts from information theory that this thesis relies on. For a more in-depth overview of these topics, see Cover & Thomas (2006). The first central concept in information theory is *entropy*, which measures the uncertainty associated with a random variable Shannon (1948). For a discrete random variable, the entropy is

[2]See Section 2.2.4 for a survey of attempts to handle apparent instances of non-compositionality under existing frameworks.

$$H(X) = -\sum_{x \in X} p(x) \log p(x). \tag{1.1}$$

Entropy is a quantity associated with a single random variable, but we are often interested in cases where there are two random variables whose outcomes affect each other. In this case, we can calculate the *conditional entropy*, which is a measure of the uncertainty associated with one of the random variables after the outcome of the other random variable is known.

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y \mid x). \tag{1.2}$$

The relationship between entropy and conditional entropy leads us to the definition of *mutual information*, which measures the change in the uncertainty of one variable that comes from knowing the outcome of the other variable. Mutual information is a symmetric, non-negative quantity and can be expressed in terms of entropy, as follows:

$$I(X;Y) = H(X) - H(X \mid Y) \tag{1.3}$$

$$I(X;Y) = H(Y) - H(Y \mid X). \tag{1.4}$$

Equivalently, it can be expressed as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{1.5}$$

It is also possible to define *pointwise mutual information* (PMI), a measure of the change in uncertainty between particular outcomes of two random variables. Unlike mutual information, PMI

can be negative (i.e., it is possible to be misinformed).

$$\text{pmi}(x;y) = \log \frac{p(x,y)}{p(x)p(y)}. \tag{1.6}$$

Mutual information takes random variables as inputs, while PMI takes outcomes of random variables. There is another family of information measures that can be described as semi-pointwise in that they take a random variable as one input and the outcome of a random variable as the other. One such semi-pointwise measure, which will be used in this thesis, is called J (Blachman, 1968). This quantity measures the amount of information that outcome y gives about random variable X. Like mutual information, J is non-negative.

$$J(X;y) = \sum_{x \in X} p(x \mid y) \log \frac{p(x \mid y)}{p(x)} \tag{1.7}$$

This thesis will also make reference to *surprisal*, which is a measure of the information content of a single outcome of a random variable:

$$\text{surprisal}(x) = -\log p(x). \tag{1.8}$$

Finally, the thesis makes reference to *relative entropy*, also known as *Kullback-Leibler distance*, which measures the distance between two probability mass functions:

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}. \tag{1.9}$$

The central (and lesser-known) information-theoretic construct used in this thesis—Partial Information Decomposition—will be introduced in Section 2.4.

1.4 A Partial Information Decomposition-based framework

The framework that I propose for measuring compositionality is rooted in information theory (Shannon, 1948; Cover & Thomas, 2006), which is concerned with quantifying the amount of information in a communication setting and is built on tools and insights from probability theory. In recent years, information theory has been used to investigate a wide range of linguistic phenomena (e.g., Goldsmith & Riggle, 2012; Smith & Levy, 2013; Futrell et al., 2019). An important characteristic of information-theoretic approaches is that they are concerned with the amount of information associated with some message rather than the content of that information.

Information theory provides a natural way of formalizing the linguistic problem at hand. Assume there is a joint distribution over meanings and linguistic forms. Let a meaning be an object $m \in M$ and let a form be an object $f \in F$. Given that we have a joint distribution between two random variables, F and M, we can condition on one or the other, and conveniently, each of the two ways of conditioning corresponds to a process in language. Conditioning on a meaning to obtain a probability distribution over linguistic forms corresponds to the process of speaking, or language production. Conditioning on a linguistic form to obtain a distribution over meanings corresponds to the process of listening, or language comprehension.

Under this setup, it is possible to talk about how informative a meaning is about a linguistic form, or vice versa. Specifically, we can talk about the mutual information between F and M, or how much information one learns about one of the random variables by observing the other.[3]

$$I(M;F) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{1.10}$$

This thesis focuses on quantifying how much of the meaning of a phrase comes from individual components of the form versus combinations of those components. Since meanings and linguistic forms both have internal structure, and since it is the relationship between subparts of these structures that we are interested in, we can define both M and F as ensemble random variables, made

[3]Mutual information allows us to quantify a relationship between two random variables, but in a linguistic context, we are more interested in how particular forms (i.e., outcomes of F) change some distribution over meanings. We can use information measure J to quantify the amount of information that individual forms give about some meaning distribution. The framework for measuring compositionality proposed in Chapter 2 uses J instead of mutual information.

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

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

Table 1.1: (left) An example of a fully systematic, or one-to-one, code, in which each variable in F is informative about a variable in M. (right) This code is less systematic because the value of each M variable depends on more than one F variable. Both codes have $I(M;F) = 2$ bits.

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 1.1. In both languages, M is an ensemble random variable made up of two binary random variables. 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.

The toy language on the left is fully systematic, whereas the toy language on the right is not, but mutual information on its own does not allow us to distinguish between these patterns. We want to be able to quantify how systematic the form-meaning mapping is, so we need a way of decomposing the mutual information such that we can identify the amount of information contributed by individual words on their own, versus the amount of information contributed by multiple words in concert. Decomposing mutual information requires extending traditional information theory to handle multivariate interactions, and P. Williams & Beer (2010)'s Partial Information Decomposition (PID) framework provides a solution to the decomposition problem. I propose that this framework is suited for our task.

P. Williams & Beer (2010) set up the problem as a decomposition of the ways that *source* variables provide information about a *target* variable. Mutual information between two source variables and one target variable can be divided up into three components: (1) the information that is provided by one source variable and no other about the target—*unique information*, (2) the information that is provided redundantly by more than one source variable—*redundant information*, and (3) the information that is provided jointly by more than one source variable, but none of them individually—*synergistic information*. I take linguistic forms (elements of F such as words) to be source variables, and meanings (elements of M) to be target variables. Under the PID framework as

applied to language, *spill* can contribute some, but not all, of its canonical meaning to the meaning of the idiom *spill the beans*, and whatever is contributed in this way is the unique information from *spill*. Meanwhile, the co-occurrence of *spill* and *beans* contributes to the idiom's meaning as well—this is synergistic information. I propose that a system in which all of the information comes non-synergistically from the source variables is a fully compositional system, and that language is subject to a principle that seeks to minimize the proportion of synergistic information, yielding a high degree of compositionality overall. Under this view, the tendency toward compositionality in language can be seen as a pressure in the system rather than a hard rule, and occurrences of non-compositionality need not be viewed as exceptional.

1.5 Thesis overview

Chapter 2 of this thesis reviews existing formalizations of compositionality, then proposes a new approach using Partial Information Decomposition. Following this are four chapters containing experimental studies looking at different facets of idiosyncrasy/context dependence in language. Note that the theory put forward in Chapter 2 was developed toward the end of the thesis process, so some of the studies reported in this thesis represent earlier stages in my thinking about compositionality and idiosyncrasy. Chapters 3 and 4, which investigate properties of idioms, use a measure called *conventionality* to capture a notion of partial compositionality. This measure can be seen as a precursor to the definition proposed in Chapter 2, which is then evaluated in Chapter 5. Chapter 6 represents an earlier use of the Partial Information Decomposition framework before I had fully developed the theory with respect to compositionality. The contributions of Chapters 3–6 are laid out in more detail below.

Chapter 3 presents experimental evidence that idioms can be characterized as occupying the intersection between two cognitive mechanisms, one that allows words to be interpreted in non-canonical ways, and one that stores linguistic structures. This is an alternative view to the popular notion of idioms as exceptions to the mechanism that builds phrases compositionally, of which there are a number of theories in the literature. An early but representative example of this position is Weinreich (1969), who posits the addition of two structures to linguistic theory: (1) an *idiom list*, where each entry contains a string of morphemes, its associated syntactic structure, and its sense

description, and (2) an *idiom comparison rule*, which matches strings against the idiom list.

In this study we define two computational measures that are intended to correspond to the two cognitive mechanisms described above. *Conventionality* is a measure of the degree to which a word has its canonical meaning in a particular context. *Contingency* measures the statistical association between words in a phrase and is intended to be a proxy for storage. In a corpus of sentences containing idiomatic and non-idiomatic phrases, we find that idioms fall at the expected intersection of low conventionality and high contingency. Additionally, we find that the head word in an idiom (e.g., the verb in a verb-object idiom) has a more conventional meaning on average than the non-head word.

Chapter 4 reports on a study of idiom production. It has been repeatedly found that idioms are processed faster than syntactically-matched literal phrases, in both comprehension and production. This has led to debate about whether idioms are accessed as chunks or built compositionally, with a series of studies in the literature attempting to measure the effect of compositionality on processing. The study in Chapter 4 looks at idiom processing through the lens of information update, in particular *surprisal theory*—a standard theory of sentence processing. The chapter argues that compositionality is just one aspect of a word's predictability, and that surprisal, as an expectation-based theory, provides a general unifying framework for understanding the idiom processing advantage. In a production experiment on verb-object idioms, we find that the idiom processing advantage can be largely explained by the fact that idioms have lower surprisal than non-idiomatic phrases. We further find that the shorter duration and lower surprisal of verb-object idioms manifest on the noun.

Chapter 5 reports on a pilot study that uses the PID-based definition of compositionality from Chapter 2 to compare the degree of compositionality between idiomatic and non-idiomatic phrases. This study proposes a way of estimating lower and upper bounds on the compositionality measure and finds that idioms come out as less compositional using this approach.

Chapter 6 extends the PID approach to compositionality to the domain of morphology. Non-compositionality and partial compositionality are not unique to idioms; the same principles apply in many other areas of language, including morphology, which is concerned with how words are built up from sub-word units like prefixes and suffixes. In some languages, individual components of a word's meaning like *singular* and *feminine* are expressed by separate affixes in a fully systematic

way, whereas in other languages, multiple units of meaning are fused together into a single affix that has no obvious decomposition (von Humboldt, 1825; Greenberg, 1960). Partial Information Decomposition provides a way of measuring variability along this dimension, known in linguistics as *fusionality*. We find that PID successfully captures the distinctions between languages with varying degrees of fusionality, providing further validation for the PID approach.

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 1.2: 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 correspondences are less common.

Chapter 7 concludes with a summary of the thesis and an overview of research questions that can be investigated using a measure of partial compositionality. The focus is primarily on questions regarding the relationship between meaning and structure, as well as how meanings change over time. I end by sketching a set of potential future studies.

Chapter 2

Formalizing partial compositionality

This chapter begins with a survey of existing attempts to formalize compositionality, along with challenges that have been raised against these attempts. Following this survey, I use information-theoretic concepts to propose a new formalism that I argue is consistent with intuitions and capable of capturing a wide range of empirical phenomena. This new formalism allows us to investigate questions that have the potential to yield new insights about the relationship between meaning and linguistic structure.

2.1 Existing frameworks

A prominent line of work in the compositionality literature has focused on providing a formal definition of the property (e.g., Montague, 1970; Hodges, 2001). The most well-known proposals are broadly similar to one another in that they consider compositionality to be a property of a meaning-form relationship as a whole—i.e., a semantics is either compositional or it is not, a binary distinction. To express the idea of compositionality mathematically, the construct of a function has typically been used to relate the meaning of an expression to the meanings of its parts and the structure combining them. Function-based approaches require a formal statement of a grammar (representing syntactic structure) as well as a formally defined semantics (conveying how meanings are represented), but there are in principle no specific requirements on how the grammar and the semantics are defined. Pagin & Westerståhl (2010a) give a useful overview of the history and details of different formalizations; I follow their general structure here.

The most influential framework comes from Montague (1970), who proposed that a compositional system is one in which there is a *homomorphism* (i.e., a structure-preserving map) between expressions in a language and their meanings. This proposal is the canonical function-based approach in the literature, and as such it requires one to define a grammar and a semantics for a language in order to evaluate whether that language is compositional. Beginning with the grammar, it is assumed that words are put together into larger phrases according to a set of syntactic rules. These rules are formalized using an algebraic structure whereby each rule is an operation defined over a set of linguistic expressions. This algebraic structure can be instantiated either as a *many-sorted* algebra, where linguistic expressions belong to syntactic categories, and each syntactic rule specifies the categories of its arguments and output (Montague, 1970; Janssen, 1986; Hendriks, 2001), or as a partial algebra, where no categories are needed, and operations may be undefined for certain arguments (Hodges, 2001). Modern discussions of compositionality generally state the formalism using partial algebras (e.g., Pagin & Westerståhl, 2010a; Szabó, 2004). Adopting this setup, we can represent the grammar (i.e., the syntax) using a partial algebra

$$\mathbf{E} = (E, A, \Sigma),$$

where E is the set of linguistic expressions in the language, A is the set of primitive linguistic atoms (e.g., words) of which E is composed, and the set Σ contains partial functions from E^n to E for each $n \geq 1$ required by the language. The set of linguistics expressions E is generated from the set of atoms A according to the rules in Σ. For a binary-branching syntax, an example of a binary partial function in Σ would be a function α that takes a determiner and a noun and returns the noun phrase as an expression:

$$\alpha(the, bird) = the\ bird$$

With the grammar in place, we move to defining the semantics. Each linguistic expression in E is associated with a derivation tree, called a *grammatical term*. Let $GT_{\mathbf{E}}$ be the set of grammatical terms for $\mathbf{E}$. The semantics is represented by a meaning function μ from $GT_{\mathbf{E}}$ to a set M of meanings. Given a grammar $\mathbf{E}$ and a semantics μ, we can express the homomorphism definition of compositionality as

(6) For every rule $\alpha \in \Sigma$ there is a meaning operation r_α such that if $\alpha(u_1, ..., u_n)$ has a meaning, $\mu(\alpha(u_1, ..., u_n)) = r_\alpha(\mu(u_1), ..., \mu(u_n))$.

While the function-based approach to formalizing compositionality is perhaps the most widely known, there is an alternative formulation that expresses nearly the same idea using the notion of substitution (Carnap, 1956; Pagin & Westerståhl, 2010a). In this setting, substitution refers to replacing elements of an expression with other elements in order to evaluate whether the meaning of the expression changes. First we define the *synonymy* relation. For $u, t \in E$,

(7) $u \equiv_\mu t$ iff $\mu(u), \mu(t)$ are both defined and $\mu(u) = \mu(t)$.

We then define the substitution version of compositionality, where the notation $s[u_1, ..., u_n]$ means that the term s contains disjoint occurrences of subterms $u_1, ..., u_n$, and that by replacing each u_i with t_i, we get $s[t_1, ..., t_n]$.

(8) If $s[u_1, ..., u_n]$ and $s[t_1, ..., t_n]$ are both meaningful terms, and if $u_i \equiv_\mu t_i$ for $1 \leq i \leq n$, then $s[u_1, ..., u_n] \equiv_\mu s[t_1, ..., t_n]$.

The substitution and function definitions of compositionality differ in whether they presuppose what is called the *domain principle*, which requires that the subterms of meaningful units be meaningful themselves. In this setup, a meaningful unit is anything in the domain of μ. The function definition presupposes this principle, while the substitution definition does not and is therefore more general. If we require the domain principle to hold, then the two definitions are equivalent.

2.1.1 Weak versus strong compositionality

There have been various proposals for both weakening and strengthening the traditional formalization of compositionality. We start by considering weaker versions, which attempt to handle the grain size issue discussed in Chapter 1. The definitions of compositionality given above state that the meaning of a linguistic expression is determined by the meanings of its *immediate* subparts and the syntax combining them. The requirement that compositionality be about immediate subparts can be weakened so that it applies to any level of the derivation. For example, a weaker version of compositionality could state that the meanings of the immediate subparts of the immediate subparts of the original expression, along with the syntactic operations at both those levels, determine the

meaning of the whole expression. This weakening can be made more extreme, with compositionality defined such that the meaning of a complex expression be determined based on the meanings of the individual atoms and the full syntactic trace of the derivation (Hodges, 2001; Pagin & Westerståhl, 2010a; Dowty, 2007). This fully weakened version does not require the meaning operation corresponding to a complex syntactic operation to have any relation to the individual operations that form its constituents (i.e., it does not enforce systematicity). It can can be expressed by a modified version of (8) where the only difference is that all of u_i and t_i are atomic.

As for strengthening compositionality, one possibility is to extend the domain of the interpretation function. Instead of compositionality making reference to the set of expressions in a particular language, one could define it such that it applies to the set of expressions across all languages (Szabó, 2000). Another way of strengthening compositionality is to introduce restrictions about the semantics, such as requiring that the interpretation function be computable. Further restrictions could then be made about what type of computable function it should be (e.g., one that operates in polynomial time).

A different type of strengthening was introduced by Hodges (2001), who posited a constraint on the substitution version of compositionality. This constraint, *inverse substitution*, says that if two terms of the same category make the same contribution to all complex expressions of which they form a part, then they have the same meaning. To express this constraint mathematically, we first assume that synonymous terms belong to the same semantic category, where we define *category* in terms of substitution: $u \sim_\mu t$ if for every term s in E, $s[u] \in dom(\mu)$ iff $s[t] \in dom(\mu)$ (Husserl, 1900; Hodges, 2001). This allows for the substitution of synonymous individual terms. We can then define inverse substitution by putting a constraint on the synonymy requirement (Pagin & Westerståhl, 2010a):

(9) If $u \not\equiv_\mu t$, then there is some term s such that either exactly one of $s[u]$ and $s[t]$ are meaningful, or both are and $s[u] \not\equiv_\mu s[t]$,

where it is assumed that synonymous terms belong to the same semantic category. The constraint can be further strengthened to disallow two expressions from being synonymous if one expression can be transformed into the other by substituting a non-synonymous term:

(10) If for some i, $0 \leq i \leq n, u_i \not\equiv t_i$, then for every term $s[u_1, ..., u_n]$ it holds that either exactly

one of $s[u_1, ..., u_n]$ and $s[t_1, ..., t_n]$ are meaningful, or both are and $s[u_1, ..., u_n] \not\equiv_\mu s[t_1, ..., t_n]$.

2.1.2 Context-dependence

The traditional framework can also be extended to handle context-dependence in language, of which two main types have been addressed in the literature. The first is extralinguistic context, which is information beyond what is contained in the utterance itself (e.g., the time and location of the utterance); for example, the word *yesterday* refers to different days depending on when it is spoken. The second type of context-dependence is concerned with the linguistic context provided by the surrounding words. There have been a few attempts to incorporate extralinguistic context into the definition of compositionality, and generally these attempts involve the incorporation of contextual arguments into the interpretation function; this can be formulated as follows, using the function definition of compositionality:

(11) For every syntactic rule $\alpha \in \Sigma$ there is a meaning operation r_α such that for every context c, if $\alpha(u_1, ..., u_n)$ has meaning in c, then $\mu(\alpha(u_1, ..., u_n), c) = r_\alpha(\mu(u_1, c), ..., \mu(u_n, c))$.[1]

When it comes to linguistic context, the definition of compositionality can be similarly extended. For this we introduce a finite set C of context types, each one corresponding to a type of linguistic context (e.g., extensional, intensional, quotation). The generalization of compositionality for this case differs from the above by allowing for each subpart of the utterance to be associated with a different context type.

(12) For every syntactic rule $\alpha \in \Sigma$ there is a meaning operation r_α such that for any context type $c \in C$ there are $c_1, ..., c_n \in C$ such that, if $\alpha(u_1, ..., u_n)$ has meaning in c, then $\mu(\alpha(u_1, ..., u_n), c) = r_\alpha(\mu(u_1, c_1), ..., \mu(u_n, c_n), c)$.

2.2 Challenges to existing frameworks

The traditional approach to formalizing compositionality has been subject to two sets of challenges. The first concerns whether the framework is mathematically suitable for distinguishing between

[1] This version of contextual compositionality entails regular compositionality, but not vice versa (Pagin & Westerståhl, 2010a).

compositional and non-compositional systems, and the second concerns whether language is in fact compositional. As we have seen, the function and substitution definitions state that compositionality is a property of a system as a whole, so as a result there have been many attempts to show that particular linguistic constructions violate the principle.

2.2.1 Conceptual and mathematical challenges

One of the main challenges to existing formalizations of compositionality holds that they are not restrictive enough. Relevant to this point are two proofs, from Janssen (1986) and Zadrozny (1994), which show that any arbitrary system can be expressed in a way that satisfies the homomorphism definition of compositionality, either by modifying the syntax or the semantics. Janssen (1986) focuses on the syntactic side, showing that if we have a meaning assignment for some recursively enumerable set of expressions, and compositionality is not satisfied, then it is possible to make the meaning assignment compositional by replacing the syntactic operations with different ones.[2] Zadrozny (1994), focusing on the semantics, shows that given a grammar, any non-compositional semantics can be rewritten (using a type-shifting operation) such that it becomes compositional and from which it is possible to recover all of the original form-meaning pairs.

These proofs are sometimes seen as an argument that compositionality is formally vacuous (e.g., Groenendijk & Stokhof, 1991; Zadrozny, 1994), though a common objection to this line of reasoning is that a grammar and a semantics should be defined based on reasonable linguistic analysis. For instance, with the Zadrozny (1994) proof, the new function that one must posit in order to make a system compositional perhaps should not be considered a meaning function at all, since it does not behave how we expect natural language semantics to behave (e.g., Szabó 2004, Kazmi & Pelletier 1998, Westerståhl 1998). As emphasized by Pagin & Westerståhl (2010a), it is always possible to enforce compositionality by unreasonable means. Janssen (1986) makes a similar point, arguing that these formal results do not mean that compositionality is a vacuous principle, and that merely proving the existence of a compositional semantics does not tell us how to obtain one. It is the work of semanticists, he argues, to come up with a plausible meaning function that satisfies compositionality.

[2]Note that there is nothing requiring the new operations to be syntactically plausible.

2.2.2 Empirical challenges

Empirical arguments against compositionality are primarily concerned with constructions in natural language that appear to violate the principle; I give an overview of some of the most commonly discussed constructions below. Many of these phenomena have received treatments in the compositionality literature showing how they can be made compatible with a compositional semantics, but in the rest of this chapter I focus on approaches to idioms in particular.

Idioms

Idioms have historically been described as exceptions to the mechanism that builds phrases compositionally, being either stored in a way similar to regular words (e.g., Swinney & Cutler, 1979), or requiring a special mechanism in the grammar (e.g., Weinreich, 1969; Bobrow & Bell, 1973). The reasoning behind these claims is that idiomatic meanings are not computable using the standard meaning composition operations for multiword phrases; for example, the idiom *kick the bucket* has the meaning "die," but this meaning is not obtainable by combining the meaning of the verb *kick* with the meaning of the direct object *the bucket* in the standard way, as it would be for a phrase like *kick the table*.

Ambiguity

In a compositional system, if there is a complex expression with a given syntactic structure and given meanings of the individual words, then it should be impossible for that expression to be ambiguous between multiple meanings, yet some sentences do seem to be ambiguous. As an example, Pelletier (2000) has argued that sentences like (13) show *quantifier scope ambiguity*. Consider the sentence below, from Pagin & Westerståhl (2010b):

(13) Every critic reviewed four films.

This sentence is ambiguous between two meanings: (1) There were four films, each of which was reviewed by every critic, and (2) Every individual critic reviewed four films, which may or may not have been the same films that the other critics reviewed. If one wants to adopt a syntactic analysis where there is a single syntactic derivation that yields both of these meanings, then this poses a

problem for compositionality.[3]

Quotation

Instances of quoted language do not behave in a standardly compositional way. As an example, the idiom *kick the bucket* is interpreted as meaning "die" in the following sentence:

(14) The old man will kick the bucket soon.

However, in a sentence talking about the idiom *kick the bucket* as a piece of language, the phrase has a different meaning.

(15) The idiom 'kick the bucket' is problematic for theories of compositionality.

This sentence does not mean "The idiom die is problematic for theories of compositionality," as it would if it were possible to transparently insert the idiomatic meaning of the phrase. For an overview of semantic treatments of quotation, see Cappelen et al. (2005).

Sentences expressing beliefs

When synonymous expressions are embedded under propositional contexts, they may no longer be synonymous. Consider this pair of sentences from Szabó (2004):

(16) a. Carla believes that eye doctors are rich.

b. Carla believes that ophthalmologists are rich.

Despite that *eye doctors* and *ophthalmologists* are synonyms, these two sentences may have different truth values if Carla does not know that the two expressions are synonymous. This violates any definition of compositionality that allows the substitution of synonymous terms.

Conditionals

Consider the following two sentences, from Szabó (2004):

(17) a. Everyone will succeed if he works hard.

[3] Alternatively, this ambiguity could be accounted for by assigning different syntactic structures to the two interpretations.

b. No one will succeed if he goofs off.

The sentence in (17a) has the interpretation $\forall x(x$ works hard $\rightarrow x$ will succeed). Assuming compositionality, we would expect the meaning of (17b) to be $\neg\exists x(x$ goofs off $\rightarrow x$ will succeed). To see why, it is useful to rewrite the predicted meaning of (17b) as follows:

(18) $\neg\exists x(x$ goofs off $\rightarrow x$ will succeed) =
$\forall x\neg(x$ goofs off $\rightarrow x$ will succeed) =
$\forall x(x$ goofs off $\wedge\neg(x$ will succeed)).

The sentence in (17b) does not mean that everyone goofs off and will not succeed. Therefore, it appears that either the quantificational noun phrases "everyone" and "no one," or the conditional "if," do not contribute information compositionally. Rather, the co-occurrence of the quantifier and the conditional contributes information necessary for computing the correct meaning of the sentence.

Anaphora

A final problem for compositionality comes from cross-sentential anaphora. Consider the following example from Szabó (2004), attributed to Barbara Partee:

(19) a. I dropped ten marbles and found all but one of them. It is probably under the sofa.
b. #I dropped ten marbles and found nine of them. It is probably under the sofa.

The first sentence of (19a) is truth-conditionally equivalent to the first sentence of (19b), and so these sentences could be considered synonymous under a semantics where truth-conditional equivalence is sufficient for synonymy. However, the second sentence in (19b), "It is probably under the sofa," is reportedly unacceptable for some speakers.

2.2.3 Test case: Idioms

This thesis focuses on idioms, which are the most well-known examples of phrases that appear to violate compositionality. Idioms are a natural area for studying contextual meaning because their idiosyncratic interpretations have been fossilized and conventionalized (i.e., stored) among a community of speakers. Note here that I assume that linguistic structures larger than words

can be stored and reused (Di Sciullo & Williams, 1987; Jackendoff, 2002), and that this storage may happen independently of any properties of the phrase's meaning (Tremblay & Baayen, 2010; O'Donnell, 2015).

Since idiomatic structures are stored and reused, we can easily locate them in corpora, obtaining information such as their frequency and the range of contexts in which they appear. This would be difficult to do with other types of phrases that have highly contextual meanings, such as novel metaphors, since there would not be multiple instances to compare (and it would be difficult to automatically detect them in a corpus).

2.2.4 Attempts to handle idioms under formal compositionality

Among those who describe idioms as exceptions to compositionality, there are two camps: one treats all idioms as non-compositional (e.g., Bobrow & Bell, 1973), while the other argues that some idioms are compositional and others are not (most prominently Nunberg et al., 1994). Common to both positions is the idea that idioms are instances of non-compositionality within an otherwise largely compositional system. Notably, this is not immediately compatible with a formal framework that treats compositionality as an all-or-nothing property of the semantics as a whole, and the framework would require significant modification in order to apply to individual phrases. There have been several attempts to rescue compositionality by modifying or constraining existing formalisms such that they handle idioms in a compositional way.[4]

Westerståhl (2002) describes three approaches to accommodating idioms under a compositional semantics, each of which has been independently proposed in the literature. These approaches are built on the homomorphism definition of compositionality, wherein idioms only constitute a problem if there does not exist a function that assigns the correct meanings to all expressions in a language, including the idioms.

The framework

All three approaches to handling idioms can be expressed using a modified version of the algebraic framework from Hodges (2001), which itself is based on the algebras for syntax and semantics

[4]There is also a significant portion of the compositionality literature devoted to handling other types of apparent non-compositionality. See Janssen & Partee (1997); Szabó (2004); Dowty (2007), among others.

initially proposed by Montague. Westerståhl (2002) describes the framework as follows.

As described in Section 2.1, we first define a grammar.

(20) **Grammar**

A grammar

$E = (E, A, \underline{\alpha})_{\alpha \in \Sigma}$

consists of a set E of **expressions**, a set $A \subseteq E$ of **atomic expressions**, and for each function symbol $\alpha \in \Sigma$ a corresponding **syntactic rule**: a partial map $\underline{\alpha}$ from E^n to E, for some n.

We then define a set of grammatical terms.

(21) **Grammatical terms**

The set $GT(E)$ of **grammatical terms** and the function $val : GT(E) \to E$ are given by:

(a) $a \in A$ is an **atomic** grammatical term, and $val(a) = a$

(b) Suppose $\alpha \in \Sigma$ is an n-ary function symbol, and $p_1, ..., p_n \in GT(E)$ with $val(p_i) = e_i$. If $\underline{\alpha}(e_1, ..., e_n)$ is defined, say $\underline{\alpha}(e_1, ..., e_n) = e$, the term $\alpha(p_1, ..., p_n)$ is in $GT(E)$, and $val(\alpha(p_1, ..., p_n)) = e$.

The grammatical terms can be thought of as derivation trees, and the function val as returning the surface string (i.e., expression) corresponding to that tree. We then define a semantics.

(22) **Semantics**

A **semantics** for E is a function μ whose domain is a subset of $GT(E)$. $p \in GT(E)$ is **μ-meaningful** if $p \in dom(\mu)$. Furthermore, p and q are **μ-synonymous**, $p \equiv_\mu q$, if $\mu(p) = \mu(q)$. $\equiv_\mu$ is an equivalence relation on $dom(\mu)$. Two semantics for μ and ν for E are **equivalent** if $\equiv_\mu$ equals $\equiv_\nu$.

This framework can be used to express different kinds of syntactic and semantic theories, so it provides a useful way of talking about the general property of compositionality, defined as follows.

(23) **Compositionality**

μ is **compositional** if $dom(\mu)$ is closed under subterms and for each $\alpha \in \Sigma$ there is a function r_α such that, whenever $\alpha(p_1, ..., p_n)$ is **μ-meaningful**, $\mu(\alpha(p_1, ..., p_n)) = r_\alpha(\mu(p_1), ..., \mu(p_n))$.

Idioms as new atoms

The simplest approach to handling idioms treats them as new atoms with corresponding idiomatic meanings; this type of analysis has been suggested by Hodges (2001) and Sag et al. (2002) for some (but not all) idioms. Under this approach, an idiom is no different than a word—indeed, one can think of it as a "word with spaces." When a phrase becomes an idiom, a new meaning is associated with the expression. In formal terms, we can say that the complex expression of interest is e_0, and we have a system with a grammar E and a semantics μ. When e_0 gains an idiomatic meaning, we add it as a new atomic expression, even though before it was a complex expression.

(24) **New Grammar**

Let $e_0 \in E - A$. The new grammar containing the idiom is $E^a = (E, A \cup \{e_0\}, \underline{\alpha})_{\alpha \in \Sigma}$

Note that no changes have been made to the syntactic rules; we have simply added an atomic expression, which is treated the same as any other expression by the existing syntactic rules. A consequence of this analysis is that subparts of an idiom cannot be targeted for syntactic transformations, and so the approach is unable to account for the fact that most idioms can undergo syntactic change. For example, consider that we can say passive sentences like "The beans were spilled," and we can add modifiers inside an idiom, as in "Leave no legal stone unturned." If we analyze the idioms *spill the beans* and *leave no stone unturned* as new atomic expressions, then we predict these transformations to be impossible (or we have to posit new atoms for every possible syntactic configuration).

Idioms as new operations

If idioms are to have internal syntactic structure, then it is still possible to handle them in a compositional semantics. We can do this by assigning idioms the same syntactic structure as their literally-interpreted counterparts. In formal terms, this can be done by using the same syntactic operation that we would use to generate the literal phrase, except we give this operation a new name and add the new version to the grammar. Thus we place a marker, or an index, on the idiomatic derivation of an expression.

(25) **New Grammar**

The grammar containing the new idiom operation is $E^i = (E, A, \underline{\alpha})_{\alpha \in \Sigma^i}$,
where $\Sigma^i = \Sigma \cup \{\alpha_0^i\}$, and α_0^i is a new k-ary function symbol such that $\underline{\alpha_0^i} = \underline{\alpha_0}$. E^i is called a **duplicated rule extension** of E. Let $q_0^i = \alpha_0^i(q_{01}, ..., q_{0k})$.

Once have adapted the grammar in this way, we can ensure compositionality in the semantics by having two different meaning operations corresponding to α and α_0^i. We extend the semantics μ such that it becomes the semantics μ^i for E^i, where $p \in dom(\mu^i) \iff p^- \in dom(\mu)$:

(26) **New Semantics**

- $\mu^i(a) = \mu(a)$ for $a \in A$ (whenever defined)
- Let $p = \alpha(p_1, ..., p_n)$ be a complex term in $GT(E^i)$. p^- is of the form $\beta(p_1^-, ..., p_n^-)$, where β is α if $\alpha \in \Sigma$, and β is α_0 if $\alpha = \alpha_0^i$. If p^- is in $dom(\mu)$ then so is each p_j^-, so $\mu^i(p_j)$ is defined, by induction hypothesis, and we let
  $\mu^i(p) = r_\alpha(\mu^i(p_1), ..., \mu^i(p_n))$ (undefined otherwise), where
  $$r_{\alpha_0^i}(m_1, ..., m_k) = \begin{cases} m_0 & \text{if } m_j = \mu(q_{0j}), 1 \leq j \leq k \\ r_{\alpha_0} & \text{otherwise} \end{cases}$$
  where each q_{0j} is a particular meaning.

This approach to maintaining compositionality allows for internal syntactic structure of idioms, but the semantics overgenerates. If we substitute a word in an idiom with a literal synonym (e.g., *kick the pail* for *kick the bucket*), then the semantics assigns the idiomatic interpretation to the substituted expression as well. That is, *kick the pail* is assigned the idiomatic meaning "die," contrary to intuitions.

Idioms as containing homophones

An alternative way of handling idioms is to treat the individual words in an idiom as new atoms, separate from their literal counterparts. By introducing homophones, we can allow the existing syntactic and semantic operations to apply. This type of approach to the semantics of idioms has been proposed by Nunberg et al. (1994) for the set of idioms that are syntactically flexible, and implemented in HPSG in Sag & Wasow (1999). First we modify the grammar in such a way that

atoms are no longer simply surface expressions; instead, the atoms in A are grammatical terms that are indexed to distinguish between homophones, and a function v turns these atoms into surface expressions in E.

(27) **New General Definition of a Grammar**

A grammar that can contain indexed atoms is

$((E, A, \underline{\alpha})_{\alpha\in\Sigma}, v)$,

where we no longer assume that $A \subseteq E$, and instead we have a function v from A to E.

Now, $a \in A$ is an **atomic** grammatical term, and $val(a) = v(a)$. To account for idioms, we add new homophonous elements $a^i_1, ..., a^i_k$ to existing atoms $a_1, ..., a_k$ whose surface forms appear in an idiom, and we assert that the new atoms are syntactically identical to the existing ones.

(28) **Idiomatic elements**

We add idiomatic elements $a^i_1, ..., a^i_k$ to A such that

$v(a^i_j) = v(a_j), 1 \leq j \leq k$.

The new grammar and semantics are as follows.

(29) **New Grammar**

The grammar containing idioms is

$E^* = ((E, A \cup \{a^i_1, ..., a^i_k\}, \underline{\alpha})_{\alpha\in\Sigma}, v^*)$,

where v^* extends v and $v^*(a^i_j) = v(a_j)$.

(30) **New Semantics**

The semantics μ^* for E^* has the following properties:

- $\mu^*(a^i_j) = m^i_j$ is added to the atomic clauses
- $\mu^*(\alpha(q_1, ..., 1_n)) = r_\alpha(\mu^*(q_1), ..., \mu^*(q_n))$
  when $\alpha(q_1, ..., q_n)^- \in dom(\mu)$.

With these adaptations in place, it is possible to build a complex expression containing the literal or the idiomatic version of a particular word. However, the changes introduced so far do

not restrict the distribution of idiomatic atoms, so the system overgenerates by allowing a sentence like *Mary tried to tie several strings* to have a meaning that involves idiomatic *strings*, contrary to intuitions. To handle this issue, further restrictions need to be introduced; one straightforward possibility is to require idiomatic words to co-occur in order for the expression containing them to be meaningful. For the idiom *pull strings*, the restriction is stated as follows Westerståhl (2002).

(31) A term in $GT(E^*)$ is meaningful iff it belongs to K^* and for each subterm of the form $\alpha(p, \alpha_0(q_1, q_2))$ it holds that $pull^i$ occurs in q_1 iff $strings^i$ occurs in either q_2 or p.

This analysis achieves better coverage than the previous two, though it encounters problems with cases where a word has its idiomatic meaning even though it does not appear with the other idiomatic words in their required positions, such as in *I would not want you to think that we are proud of our ability to pull strings, such as the ones we pulled to get you down here* (Nunberg et al., 1994; Keine, 2013). All three of the analyses described in this section rely on a clear separation between idiomatic and literal meanings. The next section will interrogate this distinction and introduce the concept of partial compositionality.

2.2.5 Partial meaning contribution

There are two ways in which a language might be considered partially compositional. The first is if the language contains some utterances that are compositional and some that are non-compositional, but for each utterance the distinction is binary. This is perhaps the more familiar notion of partial compositionality, as utterances are often described as differing in this manner. An alternative way that a language could be said to be partially compositional is if each individual utterance can obtain its meaning in a partially compositional way, with the subparts contributing variable amounts of their canonical meanings.[5] This section focuses on the latter possibility, showing how we can understand meaning contributions within individual utterances through a partial compositionality lens. Under the homomorphism-based framework, it is not coherent to describe a semantics as being partially compositional (in either sense), but I argue that adopting a notion of partial compositionality helps us better understand the empirical landscape.[6]

[5]Note that this conception of partial compositionality is agnostic to whether the varying amounts of meaning contribution are discrete or continuous.

[6]In the natural language processing literature, models of distributional semantics share similarities with the idea of partial meaning contribution, as they posit a clustering in semantic space based on the contexts in which a particular

We have seen in the previous section that an influential approach to handling idioms involves positing separate lexical items for the idiomatic versions of words (e.g., idiomatic versus literal *spill*). This approach does not capture the intuition that the amount a single word contributes to a larger meaning can vary across different phrases/contexts. In the case of *spill the beans*, for instance, this approach has no way of accounting for the fact that *spill* contributes *some* of its canonical meaning, but not as much as in *spill the coffee*. Furthermore, *spill* seems to contribute more of its literal meaning in *spill the beans* than *kick* does in *kick the bucket*.

Partial meaning contribution can also be seen with grammatical aspect, as some idioms seem to preserve the aspectual properties associated with literal uses of their components parts, displaying the full range of aspectual classes that are available for phrases of their syntactic profile. For example, the idiom *kick the bucket* cannot appear in all of the aspectual configurations that the word *die* can; it is restricted to appearing only with whatever aspect is available for transitive verbs with definite direct objects.

(32) a. Hermione was dying for weeks.

b. *Hermione was kicking the bucket for weeks. (McGinnis, 2002)

McGinnis (2002) explains the phenomenon within the framework of Distributed Morphology by appealing to a distinction between structural meaning, which is held in a word's lexical entry and is subject to syntactic operations, and idiosyncratic meaning, which applies post-syntactically and is independent of syntactic operations. She argues that structural meaning (which includes aspect) can be compositional even while idiosyncratic meaning is non-compositional.[7] However, this analysis does not account for the fact that the words in many idioms also share a portion of their idiosyncratic meaning with their literal counterparts—e.g., the fact that *spill* contributes a notion of releasing in *spill the beans*.

Partial meaning contribution is ubiquitous in language, extending far beyond idioms. It arises, for instance, in polysemy, the phenomenon whereby a word has multiple related senses, as in the following pair of sentences from Vicente & Falkum (2017). In these sentences, the word *lunch* contributes a portion of its canonical meaning in both sentences.

lexical item is used.

[7]While McGinnis (2002) describes the aspect-preservation phenomenon as being characteristic of all idioms, later work has shown that it is restricted to a subset of idiomatic phrases (Glasbey, 2007; Espinal & Mateu, 2010).

(33) a. I have my **lunch** in the backpack.

b. **Lunch** was really long today.

Unlike homophony, which is typically analyzed as involving multiple different lexical entries (e.g., *bank* as a financial institution versus *bank* of a river), polysemy has proven difficult to analyze with existing theoretical tools. One reason for this difficulty is that instances of polysemy seem to fall along a spectrum of idiosyncrasy, just as idioms do. One one end of the spectrum are instances of *regular polysemy*, which occur when the relationship between a word's multiple senses is predictable based on how other words in the language behave. Apresjan (1974) characterizes regular polysemy as occurring when, if there is a word A with senses a_i and a_j in a given language, then "there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j" (16). In English, examples of regular polysemy include words that refer to an animal or its meat, such as *chicken* and *turkey*, and artist names that can also refer to the artist's work, such as in *Proust is on the top shelf* and *Mary owns a Picasso* (Falkum & Vicente, 2015). At the other end of the spectrum, *irregular polysemy* occurs when a word's senses do not follow a pattern that exists elsewhere in the language and must therefore be memorized, as with the word *run*, which can be used in *run a mile*, *run a shop*, *run late*, and *run on gasoline* (Apresjan, 1974; Falkum & Vicente, 2015).

Regular polysemy is intuitively more compositional than irregular polysemy in the sense that it can be applied productively to new lexical items in a way that irregular polysemy cannot. Furthermore, regular polysemy is predictable based on the meaning of the individual word alone, whereas irregular polysemy relies on the specific combination of the polysemous word and the words that it combines with. There is no sharp distinction between regular and irregular polysemy (Sweetser, 1990; Taylor, 2003), as, for example, a preposition like *over* can be used in different senses that are semi-predictable (see Brugman (1988) for further discussion).

One hypothesis in early work on polysemy was that each of a word's senses is individually represented and stored in the lexicon (J. Katz, 1972; Lakoff, 1987; Brugman, 1988; Brugman & Lakoff, 1988); under such a model, homophony and polysemy are functionally equivalent. This hypothesis has been challenged for a variety of reasons. First, given that many words in a single sentence might be polysemous, and each one might have many different senses, the task of considering all

possibilities during processing could quickly become burdensome (Falkum & Vicente, 2015). Second, as argued by Sandra (1998), it is often not clear whether the interaction of a word with its context genuinely constitutes a different sense of the word. This point, which has caused a fair amount of debate in the literature (e.g., Tuggy, 1999; Gries, 2019), is the precise issue that has also caused confusion in the compositionality literature—i.e., that context-dependence is pervasive in language, and that this is in tension with semantic representations treating lexical items as dissociable units that combine in compositional, building-block fashion. Finally, there has been evidence from the psycholinguistics literature against analyzing homophony and polysemy as involving the same types of representations, with experimental results indicating that polysemy exerts a priming effect whereas homophony exerts a competition effect (Klepousniotou et al., 2008; Frisson, 2009). More recently, it has been hypothesized that polysemous words have a single representation in the mental lexicon on which each of the word's senses depends (Pustejovsky, 1995; Carston, 2012), though as far as I am aware, the details of this hypothesis have not been formally laid out (Falkum & Vicente, 2015), in large part due to the fact that existing formalisms do not allow us to talk about partial contributions of a word's meaning.

2.3 A framework for partial compositionality

As discussed in Chapter 1, linguistic meanings exhibit a great deal of context-sensitivity, while at the same time the meanings of complex expressions are highly predictable based on the meanings of their parts. In formalizing compositionality, we do not want to discount the role of either of these properties when characterizing linguistic meaning. One one hand, we want to preserve the explanatory power of systematicity in making language learnable and productive, and on the other hand, we want to preserve the intuition that expressions like idioms do in fact differ (from a compositionality standpoint) from literal phrases.

If, instead of viewing language as a completely compositional system, we treat it from a formal standpoint as a system with a pressure toward compositionality, we can preserve the relevant intuitions by allowing a spectrum of compositionality along which different constructions can fall. We can also test hypotheses about how different degrees of compositionality correlate with other properties of language, and we can characterize the full range of empirical phenomena involving par-

tial meaning contribution. In the remainder of this chapter I propose a new framework for talking about compositionality, using concepts from information theory. To begin, I submit the following desiderata for any formal characterization of compositionality:

(34) A formalization of compositionality should:

- Be able to talk about compositionality with respect to individual utterances.
- Place utterances along a scale of compositionality.
- Be able to talk about the meaning contribution of each subpart in an utterance.
- Be agnostic to how meaning is represented and therefore applicable under any semantic theory.

In seeking to satisfy these desiderata, I pursue the idea that the meaning of an utterance is made up of different types of information contributions from the individual words. Specifically, I propose a new framework for talking about compositionality, built on the following intuition:

(35) Compositionality is about how the different parts of an utterance **contribute information** to its overall meaning.

Information theory proves useful here because it allows us to quantify information contributions. We can set up the problem in the following way. Expressions in language (i.e., linguistic forms) convey some amount of information about their intended meaning, and the amount of information that is conveyed can come from different sources within the expression. For our purposes, these sources can simply be the words in the expression; for example, in the phrase *build a house*, each individual word contributes some portion of the total amount of information about the phrase's overall meaning. We can express this idea using the following notation:

$$\text{info}(w_1, ..., w_k; M) \tag{2.1}$$

where $w_1, ..., w_k$ are the words in a phrase, M is the meaning of the phrase, and info is some unspecified measure of information.

In an intuitively compositional phrase, the information about the phrase's meaning comes from the component words individually, and the information that a word contributes is consistent across any other phrases it occurs in.[8] Let us call this kind of information contribution—the information contributed by a single word regardless of its context—**unique information** U. Thus a fully literal, compositional phrase will have all of its information coming from the unique contributions of the individual words.

$$\text{info}(w_1, ..., w_k; M) = U(w_1; M) + ... + U(w_k; M) \tag{2.2}$$

As we have seen, idioms appear to violate compositionality. Consider the idiom *kick the bucket*, where the literal meanings of *kick* and *bucket* don't seem to contribute to the meaning of the phrase, and where the idiomatic meaning is only available if the specific words *kick* and *bucket* co-occur—that is, they contribute information jointly. Let us call the kind of information contribution, which requires joint knowledge of multiple words and is not obtainable from any of them in isolation, **synergistic information** S. The meaning of a fully idiomatic phrase, then, could be made up solely of the synergistic contributions of its parts.[9]

$$\text{info}(w_1, ..., w_k; M) = S(w_1, ..., w_k; M) \tag{2.3}$$

There are also phrases that do not fall straightforwardly into either category, such as *spill the beans*. In this idiom, the verb *spill* seems to contribute some of its literal meaning—namely, some notion of releasing or revealing—whereas *beans* less obviously contributes any of its literal meaning. It is possible that *beans* contributes some notion of discreteness, given that we map *beans* metaphorically onto the direct object of "reveal the information," but if so, this contribution is a small portion of the usual meaning of *beans*. Regardless of what *beans* contributes, this idiom involves some information being contributed uniquely, but also a large portion that is only obtainable by

[8] There is also a component of the phrase's meaning that comes from its syntactic structure, but for now we will assume that this contribution is constant among high and low compositionality phrases, and we will focus just on the contributions from the meanings of the phrase's subparts.

[9] In fact, there could be synergistic contributions from any subset of the component words.

knowing that the words *spill* and *beans* have co-occurred. This meaning of this idiom can therefore be said to have both a compositional contribution (the unique information) and a non-compositional contribution (the synergistic information).

$$\text{info}(w_1, ..., w_k; M) = U(w_1; M) + ... + U(w_k; M) + S(w_1, ..., w_k; M) \tag{2.4}$$

The combination of unique and synergistic information can help us make sense of many kinds of partial meaning contribution in natural language, such as the polysemy example in (33), repeated below. In both sentences, the word *lunch* contributes a portion of its canonical meaning—i.e., it contributes some unique information in both sentences. At the same time, there is another portion of each sentence's meaning that comes from knowing that *lunch* has occurred jointly with the other words in the sentence—i.e., synergistic information.

(36) a. I have my **lunch** in the backpack.

b. **Lunch** was really long today. (Vicente & Falkum, 2017)

There is a third logically possible type of information contribution, which occurs when information is contributed separately yet redundantly by more than one word in an utterance. We can call this type *redundant information*. Example of redundant information in language would be a phrases like *eat food*, where there is some portion of the total information that is conveyed by both *eat* and *food* on their own, or cognate object constructions such as *sing a song*, where the verb and noun are at least partially redundant. Unlike synergy, redundancy does not make an utterance any less intuitively compositional, since the information can be obtained from a single word.[10]

$$\text{info}(w_1, ..., w_k; M) = U(w_1; M) + ... + U(w_k; M) + S(w_1, ..., w_k; M) + R(w_1, ..., w_k; M) \tag{2.5}$$

The division of the information about the meaning of a phrase into types of information contribution is the fundamental insight behind the Partial Information Decomposition framework, first

[10]Going forward, I will abbreviate $U(w_1; M) + ... + U(w_k; M)$ as U, using subscripts when necessary to index a particular form's contribution; $S(w_1, ..., w_k; M)$ as S; and $R(w_1, ..., w_k; M)$ as R. I will only be considering cases where two linguistic forms contribute to a meaning, so there is no need to distinguish R and S of subsets of $w_1, ..., w_k$.

proposed by P. Williams & Beer (2010). In the rest of this chapter, I will use this framework to propose a definition of the *degree of compositionality* of an expression.

2.4 Partial Information Decomposition

The Partial Information Decomposition (PID) framework comes out of a branch of information theory investigating how to generalize mutual information to multivariate settings. 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:

$$I(F;M) = \sum_{f \in F} \sum_{m \in M} p(f,m) \log \frac{p(f,m)}{p(f)p(m)}. \tag{2.6}$$

When M and F are uninformative about each other, the mutual information is zero, and when they are informative, the mutual information is positive. The PID framework, first proposed by P. Williams & Beer (2010), provides a way of decomposing the quantity of mutual information into separable and interpretable parts. PID is defined for any setup in which multiple *source* random variables provide information about a *target* random variable. Define source random variables F_1 and F_2, and let them together be treated as an ensemble random variable F. Then the mutual information $I(F;M)$ can be rewritten as $I(F_1, F_2; M)$. P. Williams & Beer (2010) show that the mutual information $I(F_1, F_2; M)$ can be decomposed into the components discussed above: unique, redundant, and synergistic information. More specifically, the mutual information decomposes into the unique information of source 1 (U_1), the unique information of source 2 (U_2), the redundant information between the two sources (R), and the synergistic information of the two sources (S).

Consider a system in which two source random variables F_1 and F_2 contain information about a target random variable M. As discussed above, there are three logically possible ways that a single source variable can may convey information about the target variable. The first possibility is that the source alone provides information that is not provided by the other source. This is the *unique information*. In the Table 2.1, F_1 provides unique information about M—information that is not repeated by F_2 or reliant on F_2.

F_1	F_2	$\rightarrow M$
a	a	0
a	b	0
b	a	1
b	b	1

Table 2.1: An example of a code in which F_1 carries unique information about M.

The second possibility is that one source variable provides information that is separately—i.e., redundantly—provided by the other source variable. This is the *redundant information*. In Table 2.2, F_1 provides information redundantly with F_2 about M.

F_1	F_2	$\rightarrow M$
a	a	0
a	a	0
b	b	1
b	b	1

Table 2.2: An example of a code in which F_1 carries redundant information (with F_2) about M.

The third possibility is that the source variable is only informative when considered jointly with another source variable. This is the *synergistic information*. In Table 2.3, F_1 and F_2 provide synergistic information about M, and there is no unique or redundant information in the system.

F_1	F_2	$\rightarrow M$
a	a	0
a	b	1
b	a	1
b	b	0

Table 2.3: An example of a code in which F_1 and F_2 carry synergistic information about M.

Under the PID framework, there can in principle be any number of source variables. However,

the simplest case, and the one I will focus on in this thesis, is the case where there are two source variables.[11] In the setup where two form variables contribute information about a meaning variable, the PID framework says the following definitions hold:

$$I(F_1; M) = U_1 + R \tag{2.7}$$

$$I(F_2; M) = U_2 + R \tag{2.8}$$

$$I(F_1, F_2; M) = U_1 + U_2 + R + S \tag{2.9}$$

Equations 2.7 and 2.8 show the decomposition of the mutual information between a single form variable and the meaning variable. When there is only one form variable, there is no possibility of synergistic information, nor of unique contributions of any other form variable. Only when both sources are considered does the mutual information decompose into all four components, shown in Equation 2.9. All three equations have a mutual information term on the left, which we have a definition for, but we do not at this point know how to compute any of the terms on the right. We therefore have a system of three equations with four unknowns, which we cannot solve.

With a definition of redundant information, unique information, or the sum of the two (called union information), it is possible to solve the system of equations 2.7–2.9 for the remaining variables (P. Williams & Beer, 2010; Bertschinger et al., 2013; Gutknecht et al., 2021; Kolchinsky, 2022).[12] Much of the work in the PID literature has focused on formulating an independent definition of one of these quantities, with various options having been proposed (e.g., P. Williams & Beer, 2010; Bertschinger et al., 2014; Finn & Lizier, 2018; Makkeh et al., 2021). At present there is no single agreed-upon definition, as the suitability of a definition depends on the particular application of the

[11] Technically, the simplest case is when there is one source variable, but this case is trivial.

[12] Further details about how the solution is computed are given in Chapter 6.

framework. In the following section I argue for a particular definition of union information as being well suited to the task of measuring the compositionality of an expression.

2.5 A measure of partial compositionality

I propose the following definition of the *degree of compositionality* of an expression:

$$\text{Compositionality}(F; M) = \frac{U + R}{U + R + S} \tag{2.10}$$

This definition measures the proportion of the total information about the meaning of an expression that is contributed non-synergistically. For a multiword phrase, it is the proportion of information obtainable from the individual words alone, rather than in combination. Assuming a binary-branching syntactic structure, we can state the definition for the setup where there are two components of any non-terminal node:

$$\text{Compositionality}(F_1, F_2; M) = \frac{U_1 + U_2 + R}{U_1 + U_2 + R + S} \tag{2.11}$$

2.5.1 Choosing an information measure

When the PID framework was first proposed by P. Williams & Beer (2010), it was defined as a way of decomposing mutual information. However, more recent work has shown that other information quantities can also be decomposed using this framework—for example, Finn & Lizier (2018) and Makkeh et al. (2021) focus on decompositions of pointwise mutual information, while Kolchinsky (2022) argues that PID generalizes to any kind of information measure.

In applying PID in a natural language setting, we have two criteria for an information measure. First, we want a measure that is relative to particular words, since we are interested in how individual expressions contribute to the meanings of the phrases they occur in. To achieve this, we want to represent words not as random variables but rather as outcomes of random variables. Second, we want an information measure that is non-negative, since it has been shown in the PID literature

that a non-negative information measure guarantees that every component of the decomposition (i.e., unique, redundant, and synergistic information) will also be non-negative (e.g., P. Williams & Beer, 2010; Makkeh et al., 2021).

Fortunately, there exists an information measure that satisfies these criteria, $J(X; y)$. The measure is described in Blachman (1968) as "the amount of information that y gives about X" and can be thought of as a semi-pointwise information measure because it takes as inputs, on one side, the outcome of a random variable, and on the other side, a random variable.[13] The measure $J(X; y)$ can be expressed as follows:

$$J(X; y) = \sum_{x \in X} p(x \mid y) log \frac{p(x \mid y)}{p(x)}. \tag{2.12}$$

2.5.2 Choosing a definition

Now that we have chosen an information measure on which to perform the decomposition, we can compute the denominator of our *degree of compositionality* measure

$$\text{Compositionality} = \frac{U_1 + U_2 + R}{U_1 + U_2 + R + S} \tag{2.13}$$

since we know that

$$J(M; f_1, f_2) = U_1 + U_2 + R + S. \tag{2.14}$$

However, at this point we still do not have a way of computing the numerator since we have not determined how to compute the individual quantities in the decomposition. As we have seen, under the PID framework the following equations hold (though now we replace mutual information I with information measure J):

[13] This measure is one of two information measures discussed in Blachman (1968), both of which can be described as "the amount of information that y gives about X." Of the two measures, only $J(X; y)$ satisfies the criterion of non-negativity.

$$J(f_1, f_2; M) = U_1 + U_2 + R + S \tag{2.15}$$

$$J(f_1; M) = U_1 + R \tag{2.16}$$

$$J(f_2; M) = U_2 + R \tag{2.17}$$

Since we have a definition for J, we can compute the expressions on the left side of the equations, so we now need a definition of unique, redundant, or union information. Before suggesting a particular definition, I note that we can determine a lower and upper bound on the proposed *degree of compositionality* measure using only the equations above. This allows us to establish the range that the compositionality score should fall into, no matter which definition we choose. Notice that the sum of $J(f_1; M)$ and $J(f_2; M)$ is $U_1 + U_2 + 2R$, which is an upper bound on the numerator in our definition of compositionality (it double counts the redundancy):

$$\text{Compositionality}(f_1, f_2; M) \leq \frac{J(f_1; M) + J(f_2; M)}{J(f_1, f_2; M)} = \frac{U_1 + U_2 + 2R}{U_1 + U_2 + R + S} \tag{2.18}$$

Similarly, we have a lower bound:

$$\text{Compositionality}(f_1, f_2; M) \geq \frac{\max[J(f_1; M), J(f_2; M)]}{J(f_1, f_2; M)} = \frac{\max[U_1 + R, U_2 + R]}{U_1 + U_2 + R + S} \tag{2.19}$$

In Chapter 5 I report on a pilot study that uses estimates of these bounds on corpus data as a first step in validating the measure. As for choosing a definition, I propose that Kolchinsky (2022)'s *union information*, which defines the quantity in the numerator of the definition in (2.11), is well

suited for measuring compositionality. This definition has been shown to behave intuitively in a wide variety of settings and is well motivated in terms of set theory (see Kolchinsky (2022) for details). The definition of union information as given by Kolchinsky (2022) is

$$I_{\cup}(X_1; ..., X_n \to Y) := \inf_Q I(Q; Y) \text{ such that } \forall i X_i \sqsubset Q \tag{2.20}$$

where the random variables are ordered in some way such that $X_i \sqsubset Q$ indicates that X_i is less informative above Y than Q is. In Appendix E I discuss how union information might be approximated in a language setting, and I carry out a pilot study that involves this approximation.

2.6 Surprisal and PID

One convenient property of information measure J is that it is closely related to the information-theoretic quantity of *surprisal* and can be written as a Kullback-Leibler (KL) distance:

$$J(f_1, f_2; M) = D_{KL}(p(M \mid f_1, f_2) \parallel p(M)) \tag{2.21}$$

This KL-distance represents the change in belief about the meaning of an expression with two components that is caused by learning those two components. It turns out that the KL-distance $D_{KL}(p(M \mid f_1, f_2) \parallel p(M))$ is, under certain assumptions about language, equal to the surprisal

$$-\log p(f_1, f_2 \mid c) \tag{2.22}$$

where we assume that f_1 and f_2 are words occurring in sequence, and c is the sentential context. This relationship was described in Levy (2008); see Hoover (2024) for a discussion of the assumptions that make this equality hold. Surprisal quantifies the unexpectedness of a word (or words) in context and is another way of measuring the information update associated with a linguistic form. Surprisal is most well-known in linguistics from *surprisal theory* Hale (2001); Levy (2008); Smith & Levy

(2013); Wilcox et al. (2023), which hypothesizes that on-line sentence processing difficulty tracks the surprisal of the utterance.

Given that the information $J(f_1, f_2; M)$ is roughly equal to—and in many cases exactly equal to—surprisal, we can see it as the total amount of information update associated with f_1 and f_2 about M. Using the PID-based framework to measure compositionality therefore explicitly dissociates (1) the amount of information update associated with encountering f_1 and f_2 and (2) how that total amount of information is divided up (the decomposition into unique, redundant, and synergistic information). This makes a clear prediction: if processing difficulty solely reflects surprisal, then differences in compositionality—that is, how the total amount of information is divided up—should not be reflected in measures of on-line processing difficulty. The consequences of this prediction are explored in Chapter 4.

2.7 Comparison to previous formalisms

One of the main advantages of traditional frameworks for formalizing compositionality is that they are not tied to any particular semantic theory. As long as the semantics can be expressed as a function assigning meanings to linguistic forms, then one can evaluate whether the semantics satisfies compositionality. The PID-based approach that I have proposed preserves this property of semantic agnosticism. Decomposing an information quantity into its components makes no assumptions about how the information quantity was initially generated, nor about the relationship between the elements involved. The fact that information theory deals in amounts of information (rather than information content) makes it well-suited to the task of measuring compositionality without committing to particular meaning representations. Another central property of traditional formalizations is that they describe the state of a linguistic system as a whole, without making reference to any individual utterances within that system. In this respect, the PID-based approach differs greatly, as it allows us to capture compositionality with respect to individual utterances. PID lets us to measure varying degrees of compositionality, distinguishing between utterances that are more versus less compositional.

Finally, the definition of compositionality proposed in this thesis offers a new way of thinking about how to characterize idioms. Consider a theory of idioms like that given by Sag & Wasow

(1999), which treats the words in an idiom as separate lexical items from their literal counterparts and requires that idiomatic words co-occur. While this analysis captures the idea that all of the components of the phrase are needed for the idiomatic meaning to be available, it does not tell us how much each piece contributes. The PID framework allows us to capture this kind of variability in amount of information, as well as Nunberg et al. (1994)'s notion of the *decomposability* of idioms (except that where Nunberg et al. (1994) posited a binary distinction between decomposable and non-decomposable idioms, we can now precisely measure how decomposable particular idioms are). This precision opens up the possibility of testing predictions about the relationship between decomposability and other aspects of linguistic structure.

Preface to Chapter 3

The study reported in this chapter takes up the question of what distinguishes idiomatic versus non-idiomatic phrases. The chapter argues that idioms can be characterized as occupying the intersection of two separate phenomena: (1) words contributing non-canonical meanings in particular contexts, and (2) the storage and re-use of familiar structures. Neither phenomenon on its own is sufficient to characterize idioms, as the first also captures instances of novel metaphor while the second captures instances of fully compositional collocations. Computational measures of the two phenomena are proposed, and a corpus study shows that idioms cluster in the expected region.

The first of the two phenomena—the contribution of non-canonical word meaning to the meaning of a phrase—is closely related to the ideas about partial compositionality discussed in Chapter 2. The measure proposed here for capturing this phenomenon is a measure of the extent to which a word contributes its canonical meaning in a particular context. This measure, which I call *conventionality*, is one way of estimating the degree of contribution of a phrase's literal meaning and can be understood as a precursor to the PID-based definition proposed in Chapter 2 and implemented in Chapter 5.

Chapter 3

Characterizing idioms: Conventionality and contingency

3.1 Introduction

Idioms—expressions like *rock the boat*—bring together two phenomena which are of fundamental interest in understanding language. First, they exemplify *non-conventional word meaning* (Weinreich, 1969; Nunberg et al., 1994). The words *rock* and *boat* in this idiom seem to carry particular meanings—something like *destabilize* and *situation*, respectively—which are different from the conventional meanings of these words in other contexts. Second, unlike other kinds of non-conventional word use such as novel metaphor, there is a contingency relationship between words in an idiom (Wood, 1986; Pulman, 1993). It is the specific combination of the words *rock* and *boat* that has come to carry the idiomatic meaning. *Shake the canoe* does not have the same accepted meaning.

In the literature, most discussions of idioms make use of prototypical examples such as *rock the boat*. This obscures an important fact: There is no generally agreed-upon definition of *idiom*; phrase types such as light verb constructions (e.g., *take a walk*) and semantically transparent collocations (e.g., *now or never*) are sometimes included in the class (e.g., Palmer, 1981) and sometimes not (e.g., Cowie, 1981). This lack of homogeneity among idiomatic phrases has been recognized as a challenge in the domain of NLP, with Sag et al. (2002) suggesting that a variety of techniques are needed to deal with different kinds of multi-word expressions. What does seem clear is that

prototypical cases of idiomatic phrases tend to have higher levels of both non-conventional meaning and contingency between words.

This combination of non-conventionality and contingency has led to a number of theories that treat idioms as exceptions to the mechanisms that build phrases compositionally. These theories posit special machinery for handling idioms (e.g., Weinreich, 1969; Bobrow & Bell, 1973; Swinney & Cutler, 1979). An early but representative example of this position is Weinreich (1969), who posits the addition of two structures to linguistic theory: (1) an *idiom list*, where each entry contains a string of morphemes, its associated syntactic structure, and its sense description, and (2) an *idiom comparison rule*, which matches strings against the idiom list. Such theories must of course provide principles for addressing the difficult problem of distinguishing idioms from other instances of non-conventionality or contingency.

We propose an alternative approach, which views idioms not as exceptional, but merely the result of the interaction of two independently motivated cognitive mechanisms. The first allows words to be interpreted in non-canonical ways depending on context. The second allows for the storage and reuse of linguistic structures—not just words, but larger phrases as well (e.g., Di Sciullo & Williams, 1987; Jackendoff, 2002; O'Donnell, 2015). There is disagreement in the literature about the relationship between these two properties; some theories of representation predict that the only elements that get stored are those with non-canonical meanings (e.g., Bloomfield, 1933; Pinker & Prince, 1988), whereas others predict that storage can happen no matter what (e.g., O'Donnell, 2015; Tremblay & Baayen, 2010). We predict that, consistent with the latter set of theories, neither mechanism should depend on the other.

This paper presents evidence that prototypical idioms occupy a particular region of the space of these two mechanisms, but are not otherwise exceptional. We define two measures, *conventionality*—meant to measure the degree to which words are interpreted in a canonical way, and *contingency*—a statistical association measure meant to capture the degree to which the presence of one word form depends on the presence of another. Our implementations make use of the pre-trained language models BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019). We construct a novel corpus of English phrases typically called idioms, and show that these phrases fall at the intersection of low conventionality and high contingency, but that the two measures are not correlated and there are no clear discontinuities that separate idioms from other types of phrases.

Our experiments also reveal hitherto unnoticed asymmetries in the behavior of head and non-head words of idioms. In idioms, the dependent word (e.g., *boat* in *rock the boat*) shows greater deviation from its conventional meaning than the head.

3.2 Conventionality and contingency

In this section we describe the motivation behind our two measures and lay out our predictions about their interaction.

Our first measure, *conventionality*, captures the extent to which subparts of a phrase contribute their normal meaning to the phrase. Most of language is highly conventional; we can combine a relatively small set of units in novel ways, precisely because we can trust that those units will have similar meanings across contexts. At the same time, the linguistic system allows structures like metaphors and idioms, which use words in non-conventional ways. Our conventionality measure is intended to distinguish phrases based on how conventional the meanings of their words are.

Our second measure, *contingency*, captures how unexpectedly often a group of words occurs together in a phrase and, thus, measures the degree to which there is a statistical contingency—the presence of one or more words strongly signals the likely presence of the others. This notion of contingency has also been argued to be a critical piece of evidence used by language learners in deciding which linguistic structures to store (e.g., Hay, 2003; O'Donnell, 2015).

To aid in visualizing the space of phrase types we expect to find in language, we place our two dimensions on the axes of a 2x2 matrix, where each cell contains phrases that are either high or low on the conventionality scale, and high or low on the contingency scale. The matrix is given in Figure 3.1, with the types of phrases we expect in each cell.

	Low conv.	**High conv.**
High cont.	Idioms *(e.g., raise hell)*	Common collocations *(e.g., in and out)*
Low cont.	Novel metaphors	Regular language use *(e.g., eat peas)*

Figure 3.1: Matrix of phrase types, organized by whether they have high/low conventionality and high/low contingency

We expect our measures to place idioms primarily in the top left corner of the space. At the same time, we predict a lack of correlation between the measures and a lack of major discontinuities in the space. We take these predictions to be consistent with theories that factorize the problem into two mechanisms (captured by our dimensions of conventionality and contingency). We contend that this factorization provides a natural way of characterizing not just idioms, but also collocations and novel metaphors, alongside regular language use.

3.3 Methods

In this section, we describe the creation of our corpus of idioms and define measures of conventionality and contingency. Given that definitions of idioms differ in which phrases in our dataset count as idioms (some would include semantically transparent collocations, others would not), we do not want to commit to any particular definition a priori, while still acknowledging that people share somewhat weak but broad intuitions about idiomaticity. As we discuss below, our idiom dataset consists of phrases that have at some point been called idioms in the linguistics literature.

3.3.1 Dataset

We built a corpus of sentences containing idioms and non-idioms, all gathered from the British National Corpus (BNC; BNC Consortium, 2007), which is a 100 million word collection of written and spoken English from the late twentieth century. The corpus we construct is made up of sentences containing *target phrases* and *matched phrases*, which we detail below.

The target phrases in our corpus consist of 207 English phrasal expressions, some of which are prototypical idioms (e.g., *rock the boat*) and some of which are boundary cases that are sometimes considered idioms, such as collocations (e.g., *bits and pieces*). These expressions are divided into four categories based on their syntax: verb object (VO), adjective noun (AN), noun noun (NN), and binomial (B) expressions. Binomial expressions are fixed pairs of words joined by *and* or *or* (e.g., *wear and tear*). The phrases were selected from lists of idioms published in linguistics papers (Riehemann, 2001; Morgan & Levy, 2016; Stone, 2016; Bruening et al., 2018; Bruening, 2020; Titone et al., 2019). We added the lists to our dataset one-by-one until we had at least 30 phrases of each syntactic type. We chose these four types in advance to investigate a variety of syntactic types to

prevent our results from being too heavily skewed by any potential syntactic confounds in particular constructions. The full list of target phrases is given in Appendix A. The numerical distribution of phrases is given in Table 3.1.

Phrase type	**Number of phrases**	**Example**
VO	31	***jump*** *the gun*
NN	36	*word* ***salad***
AN	33	*red* ***tape***
B	58	***fast*** *and loose*

Table 3.1: Types, counts, and examples of target phrases in our idiom corpus, with head words bolded

The BNC was constituency parsed using the Stanford Parser (Manning et al., 2014), then Tregex (Levy & Andrew, 2006) expressions were used to find instances of each target phrase.

Matched, non-idiomatic sentences were also extracted in order to allow for direct comparison of conventionality scores for the same word in idiomatic and non-idiomatic contexts. To obtain these matches, we used Tregex to find sentences that included a phrase with the same syntactic structure as the target phrase. Each target phrase was used to obtain two sets of matched phrases: one set where the head word remained constant and one where the non-head word remained constant.[1] For example, to get head word matches of the adjective noun combination *sour grapes*, we found sentences where the lemma *grape* was modified with an adjective other than *sour*. Below is an example of a sentence found by this method:

> *Not a* ***special grape*** *for winemaking, nor*
> *a hidden architectural treasure, but hot*
> *steam gushing out of the earth.*

The number of instances of the matched phrases ranged from 29 (the number of verb object phrases with the object *logs* and a verb other than *saw*) to the tens of thousands (e.g., for verb object phrases beginning with *have*), with the majority falling in the range of a few hundred to a few thousand. Issues of sparsity were more pronounced among the target phrases, which ranged from one instance (*word salad*) to 2287 (*up and down*). Because of this sparsity, some of the analyses

[1]To obtain matched phrases, we follow work such as Gazdar (1981), Rothstein (1991), and Kayne (1994) in treating the first element in a binomial as the head. We discuss this further in Section 3.6.

described below focus on a subset of the phrases.

The syntactic consistency between the target and matched phrases is an important feature of our corpus, as it allows us to compare conventionality across semantic contexts while controlling for syntactic structure.

3.3.2 Conventionality measure

Our measure of conventionality is built on the idea that a word being used in a conventional way should have similar or related meanings across contexts, whereas a non-conventional word meaning can be idiosyncratic to particular contexts. In the case of idioms, we expect that the difference between a word's meaning in an idiom and the word's conventional meaning should be large. On the other hand, there should be little difference between the word's meaning in a non-idiom and the word's conventional meaning.

Our measure makes use of the language model BERT Devlin et al. (2019) to obtain contextualized embeddings for the words in our dataset. BERT was trained on a corpus of English text, both nonfiction and fiction, with the objectives of masked language modeling and next sentence prediction. For each of our phrases, we compute the conventionality measure separately for the head and non-head words. For each case (head and non-head), we first take the average embedding for the word across sentences *not containing* the phrase. That is, for *rock* in *rock the boat*, we get the embeddings for the word *rock* in sentences where it does not occur with the direct object *boat*. Let O be a set of instances $w_1, w_2, ..., w_n$ of a particular word used in contexts *other than* the context of the target phrase. Each instance has an embedding $u_{w_1}, u_{w_2}, ..., u_{w_n}$. The average embedding for the word among these sentences is:

$$\mu_O = \frac{1}{n} \sum_{i=1}^{n} u_{w_i} \tag{3.1}$$

We take this quantity to be a proxy for the prototypical, or conventional, meaning of the word. The conventionality score is the negative of the average distance between μ_O and the embeddings for uses of the word across instances of the phrase in question. We compute this as follows:

$$\text{conv(phrase)} = -\frac{1}{m} \sum_{i=1}^{m} \left\| \frac{T_i - \mu_O}{\sigma_O} \right\|_2 \tag{3.2}$$

where T is the embedding corresponding to a particular use of the word in the target phrase, and σ_O is the component-wise standard deviation of the set of embeddings u_{w_i}, and m is the number of sentences in which the target phrase is used.

3.3.3 Contingency measure

Our second measure, which we have termed *contingency*, refers to whether a particular set of words appears within the same phrase at an unexpectedly high rate. The measure is based on the notion of pointwise mutual information (PMI), which is a measure of the strength of association between two events. We use a generalization of PMI that extends it to sets of more than two events, allowing us to capture the association between phrases that contain more than two words.

The specific generalization of PMI that we use has at various times been called total correlation (Watanabe, 1960), multi-information (Studený & Vejnarová, 1998), and specific correlation (van de Cruys, 2011).

$$\text{cont}(x_1, x_2, ..., x_n) = \log \frac{p(x_1, x_2, ..., x_n)}{\prod_{i=1}^{n} p(x_i)} \tag{3.3}$$

For the case of three variables, we get:

$$\text{cont}(x, y, z) = \log \frac{p(x, y, z)}{p(x)p(y)p(z)} \tag{3.4}$$

To estimate the contingency of a phrase, we use word probabilities given by XLNet (Yang et al., 2019), an auto-regressive language model that gives estimates for the conditional probabilities of words given their context. Like BERT, XLNet was trained on a mix of fiction and nonfiction data. To estimate the joint probability of the words in *rock the boat* in some particular context (the numerator of the expression above), we use XLNet to obtain the product of the conditional probabilities in the chain rule decomposition of the joint. We get the relevant marginal probabilities by using attention masks over particular words, as shown below, where c refers to the context—that is, the rest of the words in the sentence containing *rock the boat*.

$$\begin{aligned}
&\Pr(\textit{boat} \mid \textit{rock the},\ c) = ..\textit{rock the}\ \textbf{boat}... \\
&\Pr(\textit{the} \mid \textit{rock},\ c) \qquad = ...\textit{rock}\ \textbf{the}\ [___]... \\
&\Pr(\textit{rock} \mid c) \qquad\quad = ...\textbf{rock}\ [___]\ [___]...
\end{aligned}$$

The denominator is the product of the probabilities of each individual word in the phrase, with both of the other words masked out:

$$\begin{aligned}
\Pr(\textit{boat} \mid c) &= ...[___]\ [___]\ \textbf{boat}... \\
\Pr(\textit{the} \mid c) &= ...[___]\ \textbf{the}\ [___]... \\
\Pr(\textit{rock} \mid c) &= ...\textbf{rock}\ [___]\ [___]...
\end{aligned}$$

The conditional probabilities were computed right to left, and included the sentence to the left and the sentence to the right of the target sentence for context. Note that in order to have an interpretable chain rule decomposition for each sequence, we calculate the XLNet-based generalized PMI for the entire string bounded by the two words of the idiom—this means, for example, that the phrase *rock the fragile boat* will return the PMI score for the entire phrase, adjective included.

3.4 Validation of conventionality measure

Our conventionality measure provides an indirect way of looking at how canonical a word's meaning is in context. In order to validate that the measure corresponds to an intuitive notion of unusual word meaning, we carried out an online experiment to see whether human judgments of conventionality correlated with our automatically-computed conventionality scores. The experimental design and results are described below. (Note that our contingency measure directly computes the statistical quantity we want, so validation is not necessary.)

3.4.1 Human rating experiment

The experiment asked participants to rate the literalness of a word or phrase in context.[2] We used twenty-two verb object target phrases and their corresponding matched phrases.[3] For each target phrase (e.g., *rock the boat*), there were ten items, each of which consisted of the target phrase used in the context of a (different) sentence. Each sentence was presented with the preceding sentence and the following sentence as context, which is the same amount of context that the automatic measure

[2]Participants were recruited on Amazon Mechanical Turk and compensated at a rate of $15/hour. The study was carried out with REB approval.

[3]We excluded one target phrase from the analyses (*spill the beans*) based on examination of the BERT-based conventionality scores. The verb *spill* used in *spill the beans* scored anomalously high on conventionality; investigation of the target and matched sentences revealed that roughly half of the matched sentences included a different idiom: *spill X's guts*. We checked the rest of our dataset and did not find other instances of this confound.

was given. In each item, a word or phrase was highlighted, and the participant was asked to rate the literalness of the highlighted element. We obtained judgments of the literalness of the head word, non-head word, and entire phrase for ten different sentences containing each target phrase.

We also obtained literalness judgments of the head word and entire phrase for phrases matched on the head of the idiom (e.g., verb object phrases with *rock* as the verb and a noun other than *boat* as the object). Similarly, we obtained literalness judgments of the non-head word and the entire phrase for phrases matched on the non-head word of the idiom (e.g., verb object phrases with *boat* as the object and a verb other than *rock*). Participants were asked to rate literalness on a scale from 1 ('Not literal at all') to 6 ('Completely literal'). We chose to use an even number of points on the scale to discourage participants from imposing a three-way partition into 'low', 'neutral', and 'high'. Items were presented using a Latin square design. The experiment was run online using the Prosodylab Experimenter (Wagner, 2021), a JavaScript tool building on jsPsych (De Leeuw, 2015).

Participants were adult native English speakers who gave written informed consent to participate. The experiment took about 10 minutes to complete. The data were recorded using anonymized participant codes, and none of the results included any identifying information. There were 150 participants total. The data from 10 of those participants were excluded due to failure to follow the instructions (assessed with catch trials).

3.4.2 Results

To explore whether our conventionality measure correlates with human judgments of literalness, we compare the scores to the results from the rating experiment. Ratings were between 1 and 6, with 6 being the highest level of conventionality.

We predicted that the literalness ratings should increase as conventionality scores increased. To assess whether our prediction was borne out, a linear mixed model was fit using the lmerTest Kuznetsova et al. (2017) package in R R Core Team (2017), with conventionality score and highlighted word (head versus non-head) and their interaction as predictors, plus random effects of participant and item.[4] All random effects were maximal up to convergence. Results are shown in Table B.1 in Appendix B. The results confirm our prediction that words that receive higher conventionality scores are rated as highly literal by humans ($\hat{\beta} = 0.185$, $SE(\hat{\beta}) = 0.050$, $p < 0.001$; see

[4] `Rating~Conv*Head+(1|Item)+(1+Conv||Partp)`

Row 2 of Table B.1 in Appendix B).

We carried out a nested model comparison to see whether including the BERT conventionality score as a predictor significantly improved the model, and we found that it did. A likelihood ratio test with the above model and one without the BERT conventionality score as a predictor yielded a higher log likelihood for the full model ($\chi^2 = 80.043$, $p < 0.001$).

3.5 Analyses

In this section we present analyses of our two measures individually, showing that they capture the properties they were intended to capture. We then investigate the interaction between the measures. Section 3.5.3 evaluates our central predictions.

We predict that the target phrases will score lower on conventionality than the matched phrases, since we expect these phrases to contain words with (often highly) unconventional meanings. We further predict that the target phrases will have higher contingency scores than the matched phrases, due to all of the target phrases being expressions that are frequently reused. Putting the two measures together, we expect idioms to fall at the intersection of low conventionality and high contingency, but not to show major discontinuities that qualitatively distinguish them from phrases that fall at other areas of intersection.

3.5.1 Analysis 1: Conventionality measure

We find that the target phrases have lower average conventionality scores than the matched phrases, with a difference of -1.654, with $t(145) = -5.829$ and $p < 0.001$. This is consistent with idioms having unconventional word meanings.

3.5.2 Analysis 2: Contingency measure

We find that, averaged across contexts, the target phrases had higher contingency scores, with a difference in value of 2.25 bits, with $t(159) = 8.807$ and $p < 0.001$.

Figure 3.2 shows boxplots of the average contingency score for each phrase type. Since many of the target phrases only occurred in a handful of sentences, we have excluded phrases for which

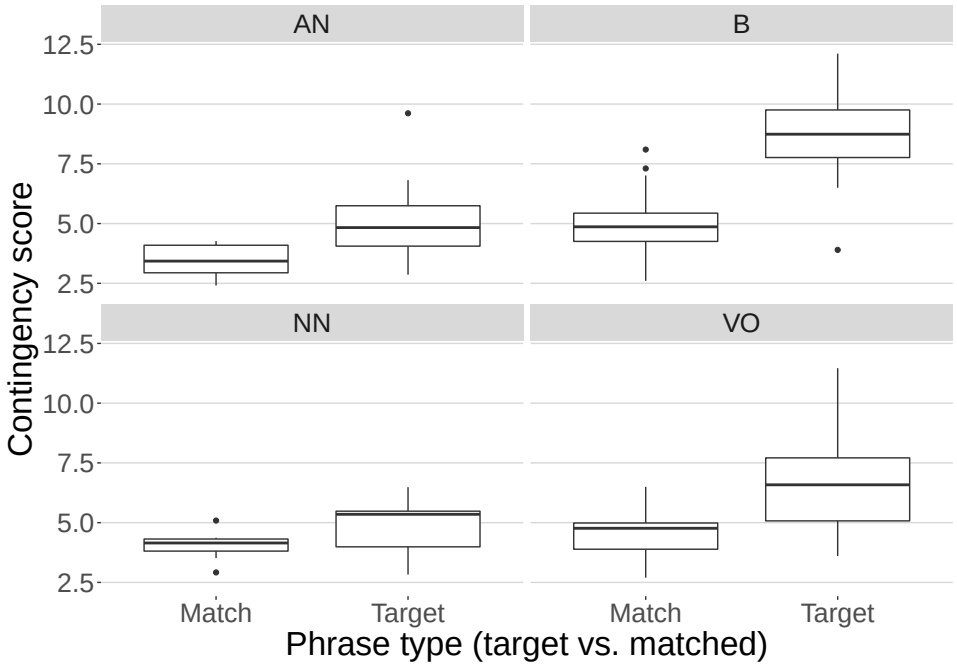

Figure 3.2: Contingency of target and matched phrases, for phrases with at least 30 instances

the target or matched sets contain fewer than 30 sentences.[5] For the most part, there were fewer sentences containing the target phrase than there were sentences containing only the head or only the non-head word in the relevant structural position. This likely explains the greater variance among the target phrases—the averages are based on fewer data points.

For all syntactic structures, the median contingency score was higher for target phrases than matched phrases. The greatest differences were observed for verb object and binomial phrases.

We fit another mixed effects model to test whether target idioms have higher contingency scores than matched phrases across syntactic classes (AN, B, NN, VO). The model predicts the contingencies for each instance of a phrase used in context, with the target-matched contrast and syntactic class as fixed effects, and random effects for the target-matched pairs.[6] We find that target phrases have significantly higher contingency scores than matched phrases (see Row 2 of Table B.2 of Appendix B).

[5]This threshold was chosen to strike a balance between having enough instances contributing to the average score for each datapoint, and having a large enough sample of phrases. We considered thresholds at every multiple of 10 until we reached one that left at least 100 datapoints remaining.

[6]`Cont~Target*Class+(1+Target|Idiom)`

3.5.3 Analysis 3: Interaction and correlation of measures

Here we show that idioms fall in the expected area of our two-dimensional space, with no evidence of correlation between the measures. Our results provide evidence against the notion of a special mechanism for idioms, whereby conventionality and contingency are expected to covary.

Recall the 2x2 matrix of contingency versus conventionality (Figure 3.1), where idioms were expected to be in the top left quadrant. Figure 3.3 shows our results. Since the conventionality scores were for individual words, we averaged the scores of the head word and the primary non-head word (i.e., the verb and the object for verb object phrases, the adjective and the noun for adjective noun phrases, the two nouns in noun noun phrases, and the two words of the same category in binomial phrases). The plot shows the average values of the target and matched phrases.

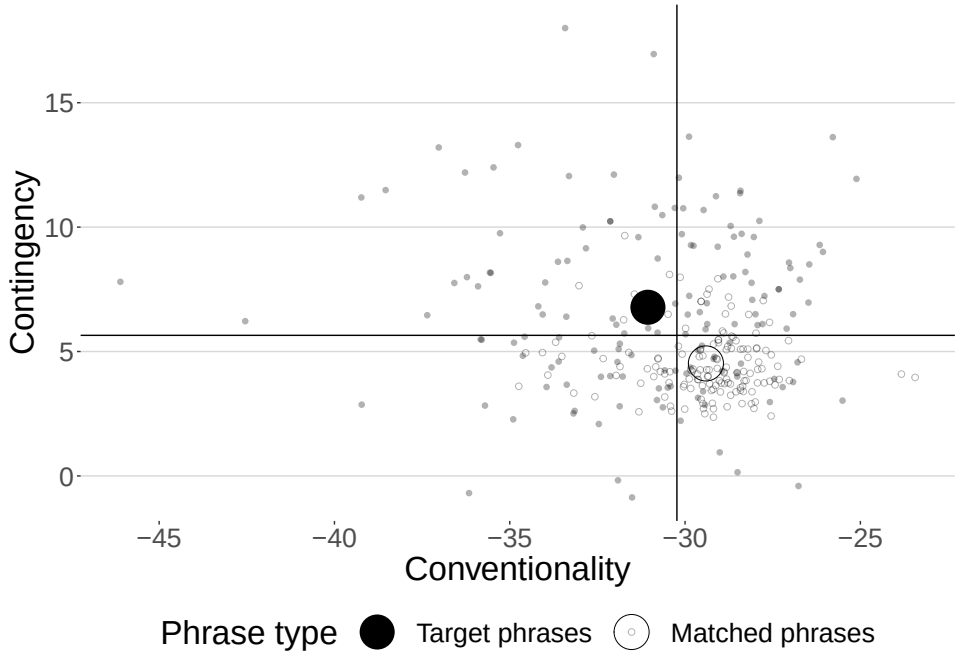

Figure 3.3: Contingency versus conventionality values of target and matched phrases. Large circles are average values of all target (black) and all matched (white) phrases.

As discussed above, the target phrases came from lists of idioms in the literature, and thus include a mix of canonical idioms and (seemingly) compositional collocations. We predicted that the target phrases would be distributed between the top two quadrants, with obvious idioms on the top left and collocations on the top right. As a sample, our results placed the following phrases in the top left quadrant: *clear the air*, *bread and butter*, *nuts and bolts*, *red tape*, and *cut corners*. For

each of these phrases, the idiomatic meaning cannot be derived by straightforwardly composing the meaning of the parts. In the top right quadrant (high conventionality, high contingency), we have *more or less*, *rise and fall*, *back and forth*, and *deliver the goods*. The bottom left quadrant was predicted to contain non-literal phrases whose words are not as strongly associated with one another as those in the most well-known idioms. The phrases in our dataset that fall into this quadrant include *hard sell*, *hit man*, and *cold feet*. A list of which target phrases landed in each quadrant is given in Appendix C.

For the matched phrases, we assumed that the majority were instances of regular language use, so we predicted them to cluster in the bottom right quadrant. Our results are consistent with this prediction. The horizontal and vertical black lines on the plot were placed at the mean values for each measure. Recall that our examples of "regular language use" consist of head-dependent constructions that share one word with an existing idiom. Although obtaining the phrases in this way may have biased our sample of "regular language use" toward similarity with target phrases, the fact that we still see a clear difference between target and matched average values is all the more striking.

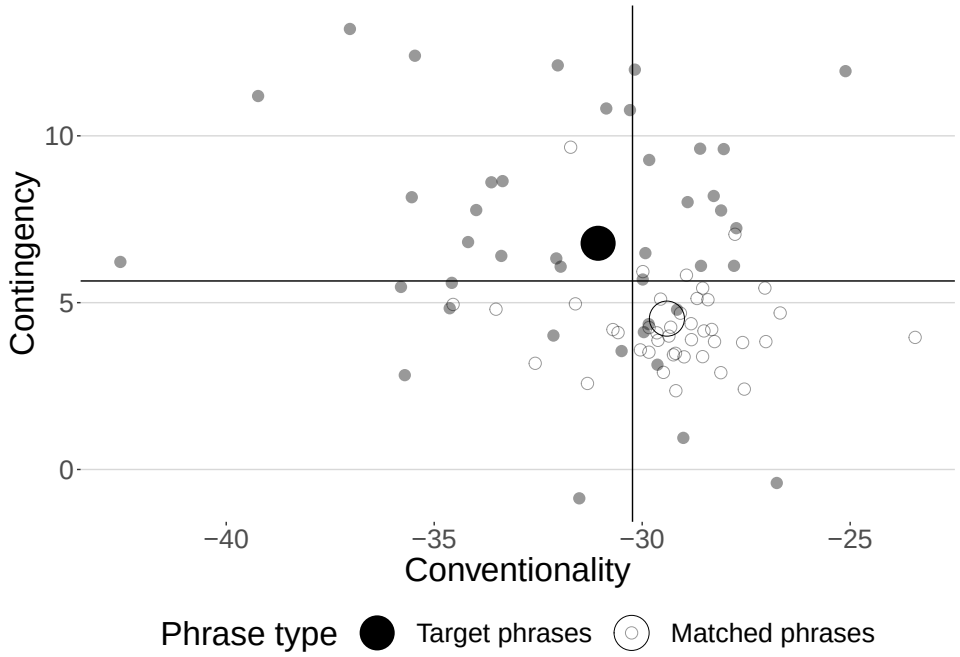

Figure 3.4: Contingency versus conventionality values of target and matched phrases (for target phrases rated as highly idiomatic). Large circles are average values of all target (black) and all matched (white) phrases.

Figure 3.4 shows only the target phrases that received a human annotation of 1 or 2 for head word literality—that is, the phrases judged to be most non-compositional. As expected, the average score for the target phrases moved more solidly into the idiom quadrant.

We also found no evidence of correlation between contingency and conventionality values among the entire set of phrases, target and matched ($r(312)$ = -0.037, $p = 0.518$), which is consistent with theories that treat the two properties as independent of each other.

3.6 Asymmetries between heads and dependents

Our experiments revealed an unexpected but interesting asymmetry between heads and their dependents. Based on conventionality scores, the head word of the target phrases was more conventional on average than the primary non-head word. A two-sample t-test revealed that this difference was significant ($t = 3.029$, $df = 252.45$, $p = 0.0027$). The matched phrases did not show a significant difference between heads and non-heads ($t = 1.506$, $df = 277.42$, $p = 0.1332$).

Figure 3.5 presents the data in a different way, with target and matched phrases plotted together. The plots show that the variability in overall phrase conventionality, which helps to distinguish idioms and non-idioms, is largely driven by the dependent word (as indicated by the steeper slopes for the non-head effects). This interaction between phrase conventionality and head/non-head is significant (see Row 10 of Table B.3 of Appendix B).

In addition, Figure 3.5 illustrates that this discrepancy between heads and non-heads is largest for verb object phrases. We confirm this by fitting a linear model of word conventionality with predictors for phrase conventionality (average of the component words), head versus non-head word, and syntactic class, plus all interactions, using sum coding to compare factor levels of syntactic class.[7] The effect of headedness on conventionality scores is significantly greater for verb object phrases than the global effect of headedness (see Panel 4 of Figure 3.5; Row 14 of Table B.3 of Appendix B). We raise the possibility that there is an additive effect of linear order, with conventionality decreasing from left to right through the phrase. For verb object phrases, the two effects go in the same direction, whereas for adjective noun and noun noun phrases, the linear order effect counteracts the headedness effect. We are not aware of any other theory positing the attri-

[7] `WordConv~PhraseConv*Class*Head`

bution of idiomatic meaning to incremental chunks in this way. Our results suggest that syntactic constituency alone is not enough to explain the observed patterns.

We note that there is disagreement in the literature about whether binomial phrases (which are coordinate structures) contain a head at all. Some proposals treat the first conjunct as the head (e.g., Rothstein, 1991; Kayne, 1994; Gazdar, 1981), while others treat the conjunction as the head or claim that there is no head (e.g., Bloomfield, 1933). We find that in the binomial case, the first conjunct patterns like the heads of the other phrase types, though how much of this effect may be driven by linear order remains unclear. This may provide suggestive converging evidence for the first-conjunct-as-head theory, though further exploration of this idea is needed.

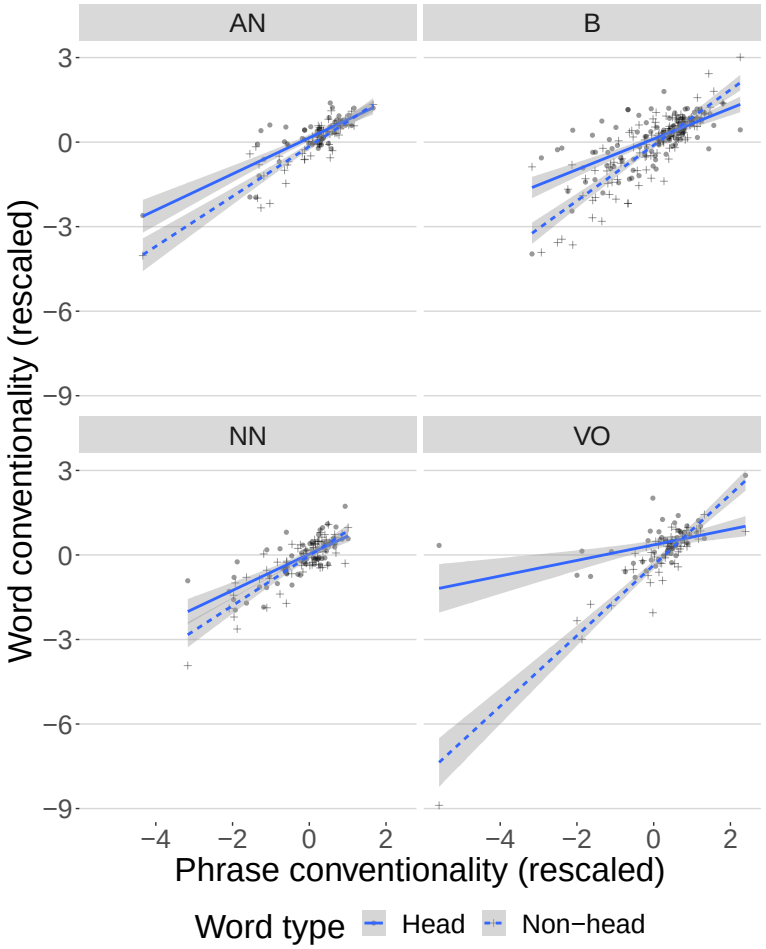

Figure 3.5: Change in head versus non-head conventionality scores as phrase conventionality increases, for all phrases (target and matched), separated by phrase type (adjective noun, binomial, noun noun, and verb object).

3.7 Related work

Many idiom detection models build on insights about unconventional meaning in metaphor. A number of approaches use distributional models, such as Kintsch (2000), Utsumi (2011), Sa-Pereira (2016), and Shutova et al. (2012), the latter of which was one of the first to implement a fully unsupervised approach for encoding relationships between words, their contexts, and their dependencies. A related line of work aims to automatically determine whether potentially idiomatic expressions are being used idiomatically or literally, based on contextual information (G. Katz & Giesbrecht, 2006; Fazly et al., 2009; Sporleder & Li, 2009, 2014). Our measure of conventionality is inspired by the insights of these models; as described in Section 3.3.2, our measure uses differences in embeddings across contexts.

Meanwhile, approaches to collocation detection have taken a probabilistic or information-theoretic approach that seeks to identify collocations using word combination probabilities. PMI is a frequently-used quantity for measuring co-occurrence probabilities (Fano, 1961; Church & Hanks, 1990). Other implementations include selectional association (Resnik, 1996), symmetric conditional probability (J. Ferreira & Pereira Lopes, 1999), and log-likelihood (Dunning, 1993; Daille, 1996). Like our study, most previous work on idiom and collocation detection focuses specifically on English.

While much of the literature in NLP recognizes that idioms share a cluster of properties, including semantic idiosyncrasy, syntactic inflexibility, and institutionalization (e.g., Sag et al., 2002; Fazly & Stevenson, 2006; Fazly et al., 2009), our approach is novel in attempting to characterize idioms along two orthogonal dimensions that correspond to specific proposals from the cognitive science literature. Our measures may offer a new avenue for tackling automatic idiom detection.

3.8 Discussion & Conclusion

We investigated whether idioms could be characterized as occupying the intersection between contingency and conventionality, without needing to appeal to idiom-specific machinery that associates the storage of multi-word expressions with the property of unconventional meaning, as has been proposed in previous work.

When we plotted conventionality and contingency scores against each other, we found that idioms fell, on average, in the area of low conventionality and high contingency, as expected. Regular, non-

idiomatic phrases fell in the high conventionality, low contingency area, also as expected. The lack of correlation between the two measures provides support for theories that divorce the notions of conventionality and contingency.

Our results suggest that idioms represent just one of the ways that conventionality and contingency can interact, analogous to collocations or metaphor. We also presented the novel finding that the locus of non-conventionality in idioms resides primarily in the dependent, rather than the head, of the phrase, a result that merits further study.

Ethics statement

This paper uses computational tools to argue for a theoretical position about idioms. Our idiom dataset was automatically generated from an existing corpus, and so did not involve data collection from human participants on our part. To validate our conventionality measure, we conducted an additional online experiment with crowdworkers on Amazon Mechanical Turk, for which we obtained REB approval. Details about the participants, recruitment, and consent process are given in Section 3.4. We note that one limitation of this work is that it only investigates English idioms, potentially contributing to an over-focus on English in this domain.

Acknowledgments

We thank Reuben Cohn-Gordon, Jacob Hoover, Alessandro Sordoni, and the Montreal Computational and Quantitative Linguistics Lab at McGill University 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.

Preface to Chapter 4

Chapter 3 introduced a measure of conventionality as a way of estimating the compositional contribution of a word to the meaning of a larger phrase. Conventionality is meant to reflect the process by which we assign non-literal interpretations to expressions and can be understood as a way of measuring how compositional a word's contribution is. Chapter 4 takes up the question of whether this measure has an observable effect on real-time sentence processing. Previous work on the processing of idioms has found that idioms are understood and produced faster than literal phrases, but has been inconclusive as to whether compositionality plays a role in this facilitation.

The study reported in this chapter investigates the roles of conventionality and surprisal in explaining the idiom processing advantage. Surprisal is a measure of the total information update associated with a phrase and is closely related to information measure J, as discussed in Section 2.6. Conventionality (which is closely related to the degree of compositionality) is orthogonal to the amount of information update. The results of our study show that processing time almost exclusively reflects surprisal, and given the close relationship between conventionality and the compositionality measure proposed in Chapter 2, the results suggest that the decomposition of the total information update into compositional versus non-compositional components plays a minimal role in explaining the idiom processing advantage.

Chapter 4

The idiom processing advantage is explained by surprisal

4.1 Introduction

Idioms have been the focus of much work in linguistics because they appear to violate a fundamental principle of human language—the principle of compositionality, which says that the meanings of individual words combine in predictable and systematic ways to yield the meanings of larger phrases (Montague, 1974; Partee, 1984; Pelletier, 2004). For example, we can understand the meaning of the phrase *eat beans* by virtue of knowing the meanings of *eat* and *beans* and the syntactic rules combining a verb with a direct object. If language were fully compositional, then we would expect an individual word to contribute a consistent meaning regardless of what other words it combines with (Wood, 1986; Pulman, 1993).

Idioms appear to violate compositionality in two ways. First, their meanings are not predictable based solely on the meanings of their component words—for instance, knowing the meanings of *spill* and *beans* is not enough to determine that *spill the beans* means something like "reveal a secret" (Weinreich, 1969; Bobrow & Bell, 1973; Swinney & Cutler, 1979; Radford, 2004; Portner, 2005). Second, idiomatic meanings are typically accessible only when specific words in a phrase occur together (thus we cannot say, e.g., *spill the legumes* and mean "reveal the secret"). There has been a large body of work, both theoretical and experimental, that aims to explain how idioms are

represented in the mind, and what this tells us about the mechanism of composition.

In early theoretical work in syntax and semantics, idioms were treated as non-decomposable lexical items; these theories are sometimes described as *words-with-spaces* theories. Within this set of approaches, some work argues that idioms are represented no differently to regular words (e.g., Swinney & Cutler, 1979), while other work argues that idioms are a special type of item with different properties than regular lexical items (e.g., Weinreich, 1969; Bobrow & Bell, 1973). However, it was later pointed out, most famously by Nunberg et al. (1994), that many idioms are in fact somewhat decomposable. Specifically, many idioms have meanings that are composed of more-or-less transparent figurative interpretations of the component words. For example, in *spill the beans*, the word *spill* contributes the figurative notion of "revealing," and *beans* is interpreted figuratively as the information being revealed. Words-with-spaces theories also encounter issues because they predict that an idiom should behave as an inseparable chunk, whereas in reality most idioms are not entirely frozen—they can often be modified (e.g., *leave no **legal** stone unturned*) and syntactically transformed (e.g., *the beans were spilled*). In more recent literature on idioms, a popular set of theories holds that the words in an idiom are actually idiom-specific lexical items that select for one another, and that they are homophonous with their non-idiomatic counterparts. For example, Sag & Wasow (1999) states that "words in...idioms take on their idiomatic meanings only when they appear together with the other parts of the idioms" (269).[1]

In parallel with theoretical work, there has been work in the psycholinguistics literature comparing the processing time of idioms and literal phrases. It has been consistently found that idioms are comprehended and produced more quickly than matched non-idiomatic phrases, especially when the idioms are well-known (e.g., van Lancker et al., 1981; Siyanova-Chanturia & Lin, 2018; Swinney & Cutler, 1979; Titone et al., 2019; Carrol & Conklin, 2020). In seeking to explain this finding, the idiom processing literature has mainly focused on the question of whether idioms are directly accessed from the lexicon during processing, as the words-with-spaces theory would predict, or whether they are built compositionally. Early research argued that idioms are directly retrieved from the lexicon as whole phrases, based on the idea that composing words into a phrase should be

[1]Furthermore, it is commonly claimed that there are different classes of idioms that are represented in different ways depending on how frozen they are (e.g., Gazdar et al., 1985; Nunberg et al., 1994; Stone, 2016); these theories typically posit that frozen idioms have a words-with-spaces representation, whereas more flexible idioms have a words-selecting-for-each-other representation.

more costly than simply retrieving a lexical item (Bobrow & Bell, 1973; Swinney & Cutler, 1979; Cacciari & Tabossi, 1988). However, more recent work has argued that a direct-retrieval story is inadequate and that meaning composition must play a role, based on experiments showing that idiom processing is sensitive to the transparency of the mapping between the idiom's meaning and the meanings of its component words, though the precise nature of the effect remains contested (e.g., Gibbs et al., 1989; Titone & Connine, 1999; Titone & Libben, 2014).

Results from the idiom processing literature have not been conclusive on the precise role of compositionality (sometimes called decomposability) on the processing time of idioms. Yet at the same time, a great deal of work in this literature assumes that compositionality is something that processing time will be sensitive to. In this paper we question this assumption; in particular, we consider another kind of theory about the factors affecting the processing time of words in context. The most well-known theory of this type is *surprisal theory* (Hale, 2001; Levy, 2008; Smith & Levy, 2013; Wilcox et al., 2023), which hypothesizes that processing time reflects how expected a word is in context. Surprisal theory holds that the processing difficulty of a word is reflected in its processing time. The processing difficulty of a word w_i is measured in *surprisal*, which is the negative log probability of the word given its context c. This quantity measures the information update associated with the word in context.

$$-\log p(w_i \mid c) \tag{4.1}$$

Previous studies of idiom processing have generally not investigated the role of surprisal in explaining the idiom processing advantage (though see Rambelli et al. 2023 and discussion in Section 4.2). We believe this is a significant oversight, as surprisal theory provides an overarching framework for capturing processing difficulty. Surprisal theory has been argued to be a *causal bottleneck* between linguistic representations and processing difficulty (Levy, 2008). In other words, surprisal subsumes various linguistic factors, and it is the surprisal quantity, rather than any of those individual factors, that is reflected in processing time. Under this framework, compositionality is merely one of many factors that goes into a word's expectedness. Surprisal is thus not an alternative to compositionality, but rather captures it. For this reason, we think that the majority

of the differences between idioms and literal phrases can be explained by the expectedness of the words in context, rather than by the decomposability of the phrase. Furthermore, we believe that the conflicting findings from previous work can be better understood if we look at them through the lens of a standard theory of processing.

Building on previous work that has reported shorter durations for idioms than for literal phrases (van Lancker et al., 1981; Siyanova-Chanturia & Lin, 2018), we investigate whether surprisal can explain this effect. Our study investigates the effects of surprisal and of *conventionality*, the latter of which was proposed in Chapter 3 as a measure of how compositional the contribution of particular words is in a larger phrases. We focus on idiom production—specifically prosody, which is known to reflect linguistic processes that may not otherwise be directly observable (F. Ferreira, 1993).

Our results confirm the finding that verb-object idioms have shorter durations than syntactically-matched literal phrases. In addition, this difference manifests on the noun—i.e, the final word of the idiom. We find that the shorter durations of idioms can be explained almost entirely by surprisal, with idiomatic nouns having very low surprisal values. This is intuitive because an idiom in context is likely to be recognized at the first word (the verb), so the noun is highly expected. Most of the literature on the idiom processing advantage is explained by this natural asymmetry with surprisal. We do also find a small conventionality effect on the verb, and we discuss possible explanations for this effect.

4.2 Relation to previous work

The only study we are aware of that comes close to investigating this question is Rambelli et al. (2023), which looks at whether compositional-but-frequent collocations pattern more like idioms or regular compositional phrases during comprehension. The authors find that collocations pattern similarly to idioms in both self-paced reading times and surprisal values.

There has been a substantial amount of work in the psycholinguistics literature on idiom comprehension, and significantly less on idiom production. Comprehension studies have overwhelmingly found that idioms are processed more quickly than literal phrase controls (e.g., Swinney & Cutler, 1979; Titone et al., 2019; Carrol & Conklin, 2020).[2]

[2]However, see Kyriacou et al. (2021) for a result that idioms are processed more slowly. The authors suggest that this effect may have been due to the cost of disambiguating between the idiomatic and literal interpretation of a

One family of predictors has been repeatedly found to speed up or otherwise facilitate idiom comprehension; these predictors seek to capture the recognizability of an idiomatic phrase (Cacciari & Tabossi, 1988; Schweigert, 1986; Cronk & Schweigert, 1992; Cronk et al., 1993; Abel, 2003; Libben & Titone, 2008; Tabossi, Fanari, & Wolf, 2009; Carrol et al., 2018; Titone et al., 2019; Carrol & Conklin, 2020). These predictors include *final-word predictability*—a measure of how likely people were, under experimental conditions, to predict the final word of an idiom given its beginning—and *familiarity*—a behavioral measure of how many people recognize a particular idiom.

Several studies on idiom processing have also included predictors related to the degree of compositionality of particular idioms. The most common such measure is *decomposability*, which comes from experimental ratings of the extent to which the words in an idiom contribute to the idiomatic meaning. There have been conflicting results regarding the effect of decomposability on idiom processing, with some studies finding a facilitative effect of decomposability (Gibbs et al., 1989; Caillies & Butcher, 2007), and others finding no such effect (Titone & Connine, 1999; Tabossi et al., 2008; Tabossi, Wolf, & Koterle, 2009; Tabossi, Fanari, & Wolf, 2009; Titone & Libben, 2014; Titone et al., 2019). Libben & Titone (2008) found a facilitative effect for offline comprehension measures (i.e., ratings and judgments), but not for an online self-paced moving window measure. Carrol & Conklin (2020) actually found that decomposability *increases* reading time. Familiarity and decomposability have also been found to interact, with Titone et al. (2019) reporting that familiarity only facilitated reading time for phrases that had low decomposability, and Libben & Titone (2008) reporting that decomposability has a greater effect for less familiar phrases.

A few studies have attempted to disentangle these results. The study reported in Titone et al. (2019), which initially looked at the effects of familiarity and decomposability on verb-object idioms using eye-tracking data, did a post-hoc analysis in which they added predictors of how related the verb and noun were to their figurative meanings. The authors found that verb relatedness made reading times faster, whereas they did not find an effect of noun relatedness. These findings suggest that the verb and noun in verb-object idioms are processed differently. In another study, Libben & Titone (2008) found that rare verbs are more predictive of verb-object idioms than common verbs, whereas the opposite is true for nouns, with more frequent nouns being more predictive of idioms than rare nouns.

phrase when preceded by a biasing context.

The studies discussed above looked at idiom comprehension; only a few studies have focused on production and prosody of idioms. Consistent with the results from comprehension studies, the consensus from work on production is that idioms are spoken faster than literal controls (van Lancker et al., 1981; Lovseth et al., 2011; Siyanova-Chanturia & Lin, 2018).

In an early study on idiom production, van Lancker et al. (1981) compared the prosody of idiomatic and literal interpretations of the same verb-object phrase. They ran two experiments, one where participants were explicitly instructed to read the sentence with a particular interpretation, and another in which participants were instructed to read the sentences naturally and were not told given a particular interpretation. The authors found that idioms were spoken faster in the experiment where participants were told to produce a particular meaning, but that there was no significant difference in the experiment where the sentences were read naturally. Lovseth et al. (2011) looked at natural productions of phrases that have both idiomatic and literal interpretations and *did* find that idiomatic phrases had shorter durations. Additionally, all of the differences the authors found between the idiomatic and literal conditions manifested on the noun. They did not find a significant effect of decomposability on duration.[3]

The goal of our study is to discover whether surprisal theory can make sense of the conflicting results above, particularly those related to the effect of compositionality on idiom processing. We suspect that a possible reason for the conflicting results about the effect of compositionality is that decomposabilityas a measure is not precise enough. Among highly decomposable idioms, the decomposability measure does not distinguish between those whose component words are semantically similar to their literal meanings (e.g., *question* in *pop the question*) and those which are less similar (e.g., *beans* in *spill the beans*). An additional problem with decomposability as a measure is that it does not capture the fact that the individual words in an idiom can differ in their conventionality (e.g., *spill* is closer to "reveal" than *beans* is to "information/secret").

In the present study, we investigate whether compositionality has any additional effect beyond that of surprisal, but instead of using decomposability as a predictor, given its drawbacks, we opt for the measure of *conventionality* from Chapter 3. Conventionality measures the distance between a word's meaning in a particular context and its canonical meaning. Socolof et al. (2022) showed that words in idioms have lower conventionality scores than the same words in syntactically-matched

[3]See Bélanger et al. (2009) for a result where idiomatic nouns were produced more slowly than literal nouns.

literal phrases, and we found this to be true in our dataset as well, for both verbs (effect size = -5.79, t(2018) = -73, $p < 0.001$) and nouns (effect size = -2.76, t(1849) = -76, $p < 0.001$). This measure addresses the issues of earlier decomposability measures because it is sensitive to *how* figurative a meaning is and because it is relativized to particular words in particular idiomatic contexts.

A significant way in which our study differs from most previous work on idiom prosody is that we compare idioms to syntactically matched phrases that do not have possible idiomatic interpretations, whereas van Lancker et al. (1981), Bélanger et al. (2009), and Lovseth et al. (2011) compared idiomatic versus literal interpretations of the same phrase. Our reasoning for this choice is to avoid any confounding effects of accessing the interpretation that is not intended in a particular trial. In addition, our study is substantially larger than the majority of studies done in this area. The following section describes our experimental setup.

4.3 Methods

We conducted an online experiment in which participants read aloud individual sentences, some of which contained idioms and some of which contained syntactically matched non-idiomatic phrases. The experiment was created using the prosodylab-Experimenter (Wagner, 2021), whose functionality is built upon jsPsych (De Leeuw, 2015).

4.3.1 Participants

We recruited 134 adult native speakers of North American English on Amazon Mechanical Turk. Of these, 44 were excluded for not being native speakers or for unintelligible recordings, leaving 90 participants (39 female, 49 male, 2 other) who ranged in age from 25 to 74 (mean 35, SD 10.9). Participants gave written informed consent and were compensated at a rate of $15/hour.

4.3.2 Materials

Each stimulus was a sentence containing a verb-object phrase. There were three conditions per item set: one where the verb-object phrase was an idiom, one where the verb-object phrase was non-idiomatic and matched the verb of the idiom, and one where the verb-object phrase was non-idiomatic and matched the noun of the idiom. The words in the verb-object phrases had the same

number of syllables across the three conditions in each item set. An example is given in (37).

(37) 1. (Idiom) The knight swore he would not rest until his enemy **bit the dust**.
2. (Verb match) The young girl blushed and **bit her lip**.
3. (Noun match) You can tell this house is old because you can really **smell the dust**.

There were 17 idioms in the experiment, each of which had 10 item sets (each consisting of one literal use of the idiom, one verb-matched literal phrase, and one noun-matched literal phrase), for a total of 510 stimuli. For each idiom, half of the item sets had the relevant phrases in sentence-final position, and half had the phrases in sentence-medial position (followed by an adjunct phrase). Every idiom in the experiment had associated conventionality scores. The item sets were split up into 10 sub-experiments, each containing one item set per idiom (so 51 stimuli). We recruited fifteen participants for each sub-experiment. Each participant saw all of the stimuli in a single sub-experiment, presented in randomized order.

4.3.3 Procedures

Participants were instructed to complete the experiment in a quiet room. They first performed a microphone check, during which they recorded themselves saying a sentence out loud with their computer microphone and listened back to the recording. At the beginning of the experiment, participants read an instruction screen, where they were told that they would be presented with one sentence at a time, and that they should first read the sentence silently to themselves, then click the record button and read the sentence out loud, and finally click the button to stop recording. They were told to read the sentences as naturally as possible. Once consent was given and the experiment began, the trials appeared in full-screen mode. For each trial, the sentence to be read aloud was written in boldface in the center of the screen. Below the sentence was text that read, "Read the sentence silently to yourself. When you're ready, click the button below to record." Below this was a button that read, "Click here to start recording." After the participant clicked this button, they were taken to a screen with the same sentence in bold in the middle of the page, with the text, "Please speak now," as well as a button reading, "Click here when you're done recording." Clicking this button moved them to the next trial. Participants could choose to press the space bar instead of clicking the button if they wished.

4.3.4 Acoustic analysis

Each sentence recording was force aligned using the Montreal Forced Aligner (McAuliffe et al., 2017). There were 4860 individual recordings from the participants who successfully followed instructions on the practice sentence. Of these recordings, 4794 were successfully force-aligned to the sentence in question. Acoustic measures of duration, pitch (mean, max, and min), and intensity (mean, max, and min) were then extracted for the words of interest (the verb and noun in the verb-object phrase, plus the first non-function word in the sentence and the sentence-final word for those stimuli where the verb-object phrase was not sentence-final).[4]

4.3.5 Surprisal and conventionality values

For each stimulus sentence in the experiment, surprisal values were estimated using language model GPT-3 for each word in the target phrase given the preceding words. Conventionality scores for the target verb and noun in each sentence were computed using the procedure described in Socolof et al. (2022), which uses the BERT language model Devlin et al. (2019) to measure the distance between a word's embedding in a particular context and the word's average embedding across other contexts.

4.4 Results

We first report the difference in duration between idiomatic and non-idiomatic phrases as a whole, then we report statistical models investigating the effects of surprisal and conventionality on verbs and nouns separately. When reporting results, we use average phone duration as our duration measure, and we compute whole phrase duration as the sum of verb and noun duration (recall that all conditions were matched for number of syllables as well as presence of a determiner between the verb and the noun). We find that idioms were produced with significantly shorter durations than literal phrases, in line with most of the existing literature. This effect is notably quite small; in our experiment, there was an average difference of 9 ms between idioms and literal phrases.

Figures 4.1 and 4.2 show verb and noun durations, respectively, for idioms versus literal phrases. The plots show that the noun is shortened in the idiom condition, whereas there is no obvious

[4]Duration was the only measure for which we saw any effects, so we focus on this measure in our results.

difference in verb duration between idioms and literal phrases.

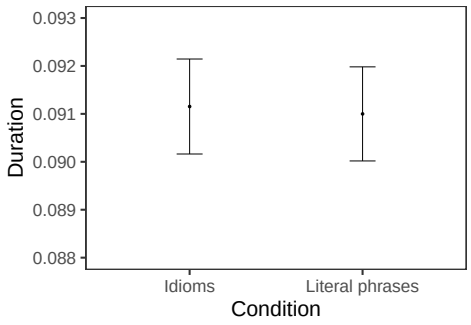

Figure 4.1: Duration of verb in verb-object idioms and verb-matched literal phrases

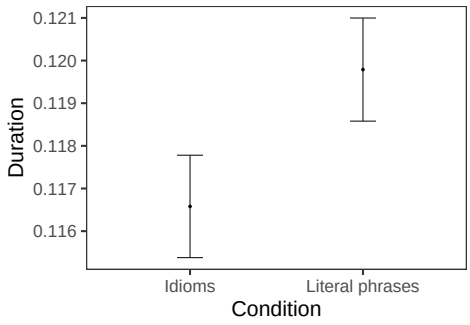

Figure 4.2: Duration of noun in verb-object idioms and noun-matched literal phrases

Looking now at surprisal, we find that idiomatic and literal phrases differ significantly in overall surprisal, with idiomatic phrases having lower surprisal (mean = 9.41) than literal phrases (mean = 11.97). This difference was driven primarily by the noun, with the nouns in idiomatic verb-object phrases having an average surprisal of 2.93 bits, whereas the nouns in literal phrases had an average surprisal of 6.06 bits. There was a much smaller difference between verbs across conditions, and the effect actually went in the opposite direction, with idiomatic verbs having higher surprisal (mean = 6.48) than literal verbs (mean = 5.87).[5] Figures 4.3 and 4.4 show the differences in surprisal between idioms and literal phrases for the verb and the noun, respectively.

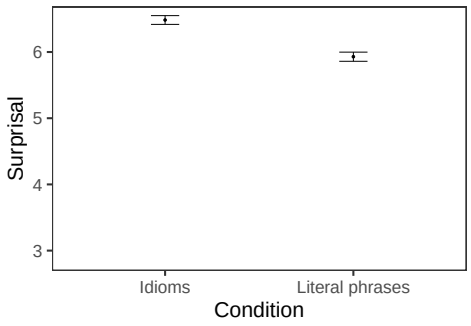

Figure 4.3: Duration of verb in verb-object idioms and verb-matched literal phrases

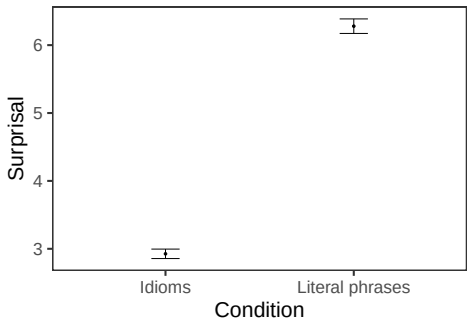

Figure 4.4: Duration of noun in verb-object idioms and noun-matched literal phrases

[5]Note that considering verb and noun surprisal together reveals that idiomatic phrases have less surprisal overall than matched literal phrases.

We now turn to the results for the conventionality measure. Conventionality, when computed for idioms, is meant to capture how different the meaning of a word is in idiomatic contexts from the word's literal meaning, and it is measured for individual words rather than full phrases.

To investigate the effects of surprisal and conventionality on the spoken duration of idioms versus literal verb-object phrases, we ran separate models for the verb and the noun. We begin with the model for verb duration. We fit a linear mixed model with verb duration as the dependent variable and verb surprisal and verb conventionality score as main predictors, as well as the surprisal of the preceding word (to account for spillover effects). Model comparison indicated that including the interaction of verb surprisal and verb conventionality improved the model. We also included random effects of item, participant, and target phrase. The model was fit using the `lmerTest` (Kuznetsova et al., 2017) package in R (R Core Team, 2017), with the formula

(38) `verbDuration ~ verbSurprisal + verbConventionality + previousWordConventionality + verbSurprisal:verbConventionality + MedialOrFinal (1 | itemNumber) + (1 | participant) + (1 | phrase).`

The continuous variables were rescaled by centering and dividing by two standard deviations. Results are shown in Table 4.1.

Table 4.1: Model results table with verb duration as the dependent variable

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
Intercept	0.075	0.051	1.45	0.151
VerbSurprisal	0.125	0.022	5.70	< 0.001*
VerbConventionality	-0.064	0.017	-3.72	< 0.001*
PreviousWordSurprisal	0.132	0.021	0.624	< 0.001*
MedialOrFinal	-0.111	0.052	-2.11	0.037*
VerbSurprisal:VerbConventionality	0.171	0.042	4.05	< 0.001*

$n = 2598$

We find that verb surprisal, previous word surprisal, and the interaction between verb surprisal and verb conventionality have the largest effects on verb duration, and that there is an additional, smaller but still significant, main effect of verb conventionality. The main effects indicate that (1) surprising verbs are spoken more slowly than predictable verbs, and (2) verbs with unconventional meanings in context are spoken more slowly than more literal usages of verbs. The effect sizes indicate that changing one unit of surprisal has an effect more than twice as big as the effect induced

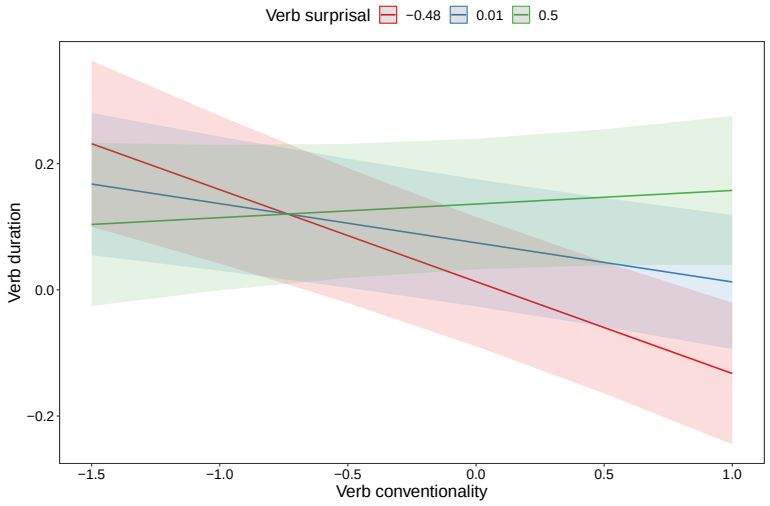

Figure 4.5: Interaction of verb surprisal and verb conventionality on verb duration

by changing one unit of conventionality. As for the interaction of surprisal and conventionality, we find that for surprising verbs, conventionality has no significant correlation with duration, whereas for predictable verbs, greater conventionality is correlated with shorter duration. The interaction is shown in Figure 4.5.

Next, we fit a model for noun duration. Since the noun is encountered after the verb, we included verb-related predictors, choosing from among models using a nested likelihood-ratio test. We found that the best model did not include verb conventionality. Our model is given by the following formula:

(39) `nounDuration ~ nounSurprisal + nounConventionality + nounSurprisal:nounConventionality + verbSurprisal + nounSurprisal:verbSurprisal + MedialOrFinal + (1 | itemNumber) + (1 | participant) + (1 | phrase).`

Again, all continuous predictors were rescaled by centering and dividing by two standard deviations. Results are shown in Table 4.2.

We find that noun surprisal has a large effect on the duration of the noun, with more surprising nouns being pronounced more slowly. Unlike for verbs, we do not find any additional significant additional effect of noun conventionality on noun duration. There was no significant interaction between noun duration and noun conventionality. We additionally find that the surprisal of the verb was positively correlated with the duration of the noun, and that there was an interaction between noun surprisal and verb surprisal, whereby the slowdown associated with surprising nouns

Table 4.2: Model results table with noun duration as the dependent variable

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
Intercept	0.236	0.069	3.43	0.002*
NounSurprisal	0.234	0.022	10.85	< 0.001*
NounConventionality	-0.031	0.019	-1.69	0.092
VerbSurprisal	0.115	0.018	6.49	< 0.001*
MedialOrFinal	-0.373	0.056	-6.67	< 0.001*
NounSurprisal:NounConventionality	-0.023	0.037	-0.63	0.532
NounSurprisal:VerbSurprisal	-0.132	0.034	-3.93	< 0.001*

$n = 2674$

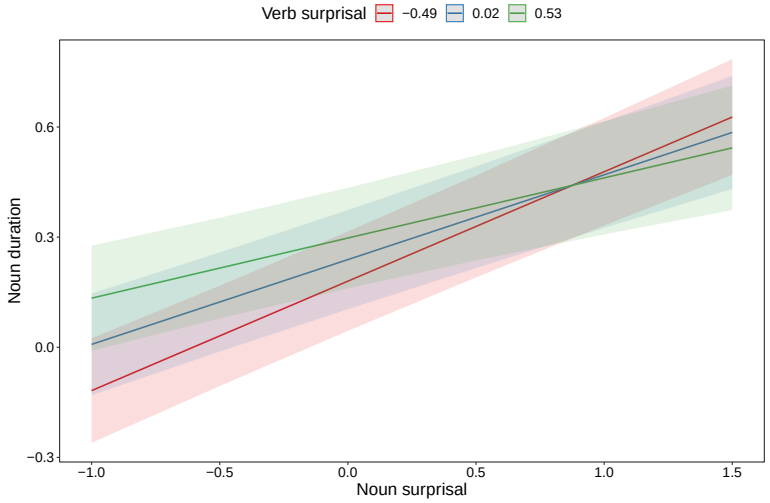

Figure 4.6: Interaction of noun surprisal and verb surprisal on noun duration

was more pronounced when the preceding verb was highly predictable. The interaction is plotted in Figure 4.6. Finally, we saw a large effect of sentence-final lengthening on the noun, which we take to be orthogonal to to the research questions at hand.

4.5 Discussion

Our first finding is that verb-object idioms are spoken more quickly on average than literal verb-object phrases, which is consistent with the consensus finding in previous work. We further find that the overall shorter duration of idioms is due to a difference between the nouns in idiomatic versus literal phrases, and that there is no significant difference between the duration of verbs across the two groups. The main aim of our experiment was to see whether this difference could be explained in

terms of surprisal—that is, the amount of information update associated with the word—as surprisal theory predicts. We find that the shorter duration of idioms can indeed be attributed in large part to the low surprisals associated with idiomatic nouns, which provides a simple explanation for the idiom processing advantage. We only find a small effect of conventionality on duration, and only on the verb. This effect is less than half the size of surprisal on verb duration.

What do the results of our models suggest about the mechanisms underlying idiom processing? We begin by considering our findings in light of the fact that processing a written phrase is a sequential phenomenon, proceeding left to right for English sentences, and we assume that people generate predictions about the next word(s) as they go along. In our stimuli, the participant first reads the sentential context up to the point where the target verb-object phrase begins, and we assume that the participant at this point is considering a distribution over next words based on the context that came before. In this distribution, there is a large amount of probability mass on individual words whose canonical meanings make sense given the context. At the same time, there is likely *also* a sizeable amount of probability mass on the first word of an idiom whose overall meaning makes sense in the context.

In the verb model, we see that high surprisal causes a slowdown, in line with surprisal theory. At the same time, the largest effect comes from the interaction of verb surprisal and verb conventionality, where the correlation between verb conventionality and verb duration changes direction depending on the surprisal of the verb. When a verb has low surprisal (i.e., it is expected given the preceding context), then unconventionality of the verb meaning is associated with longer duration, which we take to mean longer processing time.[6] But when a verb is very surprising given its preceding context, then unconventional meaning does not have a significant effect on processing time.

We posit that this pattern can be explained in the following way: there is a processing cost associated with the point at which one computes the meaning of a word and integrates it with the (incremental) meaning of the sentence, but this computation does not necessarily take place at the moment the word is encountered. We explore the different possible cases below.

Among unsurprising verbs, highly conventional ones are processed the fastest, likely because

[6] Alternatively, we can think of production time in terms of the speaker providing information to a listener, where a slowdown at a particular word indicates that that word is highly informative and should be paid attention to.

they were expected based on the preceding words, and so the person has already updated their beliefs about the meaning of the sentence with this word accounted for. If a verb is unconventional but still highly predictable, that indicates that an idiom is highly probable in that context, and the slowdown comes from computing the meaning of the full idiom, even though only the verb has been encountered.

Now let us consider surprising verbs. A verb that is surprising will cause a slowdown regardless of its conventionality, and the person may access multiple possible meanings, both conventional and unconventional, in order to figure out which one was intended. Alternatively, they may simply wait for subsequent words before trying too hard to figure out the verb's meaning, which would be compatible with there being no significant difference between conventional and unconventional verbs. In this situation, the word(s) that come after the verb will be expected to provide additional information beyond what the verb contained.

Moving now to the noun, we find a very large effect of surprisal, with low surprisal being associated with shorter duration. Furthermore, we find that low surprisal is associated with idiomatic nouns, which goes a long way toward explaining the overall idiom processing advantage. When one has already encountered a verb that strongly predicts a particular idiom, then seeing the noun that completes that idiom will simply confirm one's expectation. In addition, literal phrases have higher surprisal overall than idioms, indicating that when there is an idiom, there is less overall information to integrate into one's beliefs about the meaning of the sentence. The fact that verb surprisal and verb duration have a small lengthening effect on the noun can be attributed to a spillover effect of the processing cost of the verb. In sum, we found that the idiom processing advantage is driven by the noun in verb-object idioms, and that the bulk of the effect can be explained by surprisal without making reference to compositionality, which helps shed light on why studies that probed for compositionality without taking surprisal into account have had mixed results.

A remaining question is why we see this residue of conventionality affecting the duration of the verb at all. If surprisal were a true causal bottleneck, then we would not expect to see any residue, so the fact that we do could raises a few possibilities. One is that our surprisal estimates are simply imperfect, due to being computed from corpus statistics, and therefore do not quite capture everything they theoretically should. Alternatively, it could be that there really is some sentence processing work that surprisal does not capture, in which case the causal bottleneck is not total.

If this is the case, then the reason we only see the effect on the verb might be that most of the work of processing an idiom takes place on the verb (since the noun is so predictable), or it could be that the noun is produced so quickly that we are simply seeing a floor effect, where the effect of conventionality is present but not detectable.

4.6 Conclusion

We investigated the differences in spoken duration between verb-object idioms and syntactically-matched non-idioms, as well as the effects of the surprisal and conventionality on duration. We found that idioms are produced more quickly than non-idioms, and that this difference manifests on the noun and not the verb. We also found that surprisal and conventionality interact to affect verb duration, but that noun duration primarily reflects surprisal. An important takeaway from our study is that there is an asymmetry between verbs and their objects in whether conventionality affects their processing time.

Our results raise the question of whether a similar pattern exists in phrases with syntactic structures other than verb-object. Socolof et al. (2022) found that across syntactic structures, head words tend to have more conventional meanings than their dependents. It would be interesting to investigate the prosody of idioms where the syntactic head of the phrase comes after the dependent, as this would help tease apart effects of left-to-right processing versus head-dependent asymmetries on production. For example, what does the pattern look like for adjective-noun idioms such as *cold feet* and *sour grapes*? These kinds of studies will be necessary for a full understanding of the prosodic signature of idioms.

Preface to Chapter 5

While the previous two chapters relied on the conventionality measure as an approximation of compositionality, Chapter 5 reports on a pilot study aimed at validating the *degree of compositionality* measure proposed in Chapter 2. The degree of compositionality measure is couched in an information-theoretic framework that allows us to refine the notion of what it means to contribute meaning compositionally, and the study reported here tests this measure on a subset of the corpus created for the study in Chapter 3. The work described in this chapter represents a first step at operationalizing a theory of partial compositionality and was completed at the end of the thesis process. I hope to develop these methods and perform a more robust set of evaluations in future work.

Chapter 5

Measuring idiom compositionality using Partial Information Decomposition

5.1 Introduction

This chapter reports on a pilot study evaluating the *degree of compositionality* measure proposed in Chapter 2. The measure captures the central intuition behind the principle of compositionality—that it describes the phenomenon whereby the meaning of a complex expression is built in a predictable, systematic way from the meanings of its parts—but differs from existing formalisms in that it is relative to individual expressions and can talk about the amount of compositionality within a particular expression. The proposed measure uses the framework of Partial Information Decomposition, which provides a way of mathematically distinguishing between the compositional and non-compositional (i.e., context-dependent) portions of the meaning of an expression. The study in this chapter compares bounds on the compositionality scores for a set of idiomatic phrases and a set of literal phrases in a corpus of naturally-occurring spoken and written language, showing that one possible implementation provides supportive evidence that the measure can distinguish between high and low compositionality phrases. The particular implementation-level decisions made are fairly coarse and could be refined in a variety of ways (which will be discussed at the end of chapter), but the fact that even a coarse implementation of a pilot experiment yields results in the predicted direction provides evidence in favor of the PID-based measure.

As discussed in Chapter 2, calculating the degree of compositionality of a phrase involves decomposing the total amount of information that the components (e.g., words) $f_1, ..., f_n$ convey about the meaning M of the phrase. I use information measure J as the quantity to be decomposed, as it satisfies the criteria of being pointwise with respect to individual words and being non-negative. As discussed in Section 2.5.1, J is a KL divergence that measures how much a distribution M over meanings changes once some linguistic expressions $f_1, ..., f_n$ are encountered.

$$J(M; f_1, ...f_n) = \sum_m p(m \mid f_1, ..., f_n) \log \frac{p(m \mid f_1, ..., f_n)}{p(m)} \tag{5.1}$$

This study focuses on verb-object phrases, where the set of words of interest $f_1, ..., f_n$ consists of a verb and its direct object noun.[1] The information measure we are interested in is therefore the information J that the verb v and noun n convey about the phrase's meaning M.

$$J(M; v, n) = \sum_m p(m \mid v, n) \log \frac{p(m \mid v, n)}{p(m)} \tag{5.2}$$

Partial Information Decomposition allows us to decompose the quantity $J(M; v, n)$ into the unique, redundant, and synergistic contributions of v and n. The degree of compositionality measure is the non-synergistic proportion of the total information:

$$\text{Compositionality}(v, n; M) = \frac{U_v + U_n + R}{U_v + U_n + R + S} \tag{5.3}$$

This study compares the range between lower and upper bounds on the compositionality score for idioms versus non-idioms, as well as between idioms. The analysis is conservative in that it does not require choosing among the various definitions that have been proposed for decomposing information measure J, but rather considers bounds on the measure. The following section will lay out the quantities to be computed.[2]

[1] Some of the phrases in this study also include a determiner as part of the direct object, but given that these are function words, I leave them out of the computation. It is of course possible in principle to include them, either by first computing the compositionality of the determiner+noun constituent, then doing the same for the verb+direct object constituent, or by letting all three words be sources simultaneously.

[2] For a version of the experiment that approximates the compositionality measure itself rather than bounds, see Appendix E.

5.2 The compositionality score range

The PID framework assumes that the following equations hold:

$$J(v; M) = U_v + R \tag{5.4}$$

$$J(n; M) = U_n + R \tag{5.5}$$

$$J(v, n; M) = U_v + U_n + R + S \tag{5.6}$$

Since we have a definition for J, we can compute the expressions on the left side of the equations, yet the system of equations remains underdetermined since there are four unknowns. In order to decompose J into its unique, redundant, and synergistic components, we need a definition of one of these quantities that allows us to solve for the rest—specifically, a definition of unique information, redundant information, or their sum (union information). However, we can define bounds on the quantity we want. The sum of $J(v; M)$ and $J(n; M)$ is an upper bound for the numerator in our definition of compositionality (it double counts the redundancy).

$$\text{Compositionality}(v, n; M) \leq \frac{J(v; M) + J(n; M)}{J(v, n; M)} = \frac{U_v + U_n + 2R}{U_v + U_n + R + S} \tag{5.7}$$

The lower bound is the maximum of $J(v; M)$ and $J(n; M)$ (it omits the unique contribution of one of the words):

$$\text{Compositionality}(v, n; M) \geq \frac{\max[J(v; M), J(n; M)]}{J(v, n; M)} = \frac{\max[U_v + R, U_n + R]}{U_v + U_n + R + S} \tag{5.8}$$

I use the term *compositionality range* to refer to the range between these bounds. The objective of this study is to compare the compositionality ranges between idiomatic and literal verb-object phrases, with the prediction that the range for idioms will be shifted higher idioms than for literal phrases. If the upper and lower bounds on the degree of compositionality confirm the prediction by being lower for idioms than for literal phrases, then that is suggestive evidence in favor of the PID-based definition of compositionality.

5.3 Methods

5.3.1 Corpus

The corpus for this pilot experiment was generated using the same procedure as in the study in Chapter 3. As described in Chapter 3, sentences containing idiomatic and non-idiomatic phrases were extracted from the British National Corpus (BNC; BNC Consortium, 2007), a 100 million word collection of written and spoken English from the late twentieth century. Eighteen verb-object idioms were selected for this experiment. To be selected, an idiom had to occur at least 10 times in the BNC, and also had to have the potential for syntactically-matched literal phrases—that is, the verb had to allow a variety of direct objects in literal usages. Idioms that were excluded due to this second criterion were those in which the verb, when used in a verb-object structure, was almost always idiomatic, such as in *talk turkey* and *turn tail*. With *talk turkey*, for example, syntactically-matched phrases also tended to be idiomatic (e.g., *talk shop*). For each idiom in the set of 18, all sentences containing the idiom were pulled from the BNC, as were all syntactically-matched phrases that shared the same verb.

The BNC was constituency parsed using the Stanford Parser (Manning et al., 2014), then Tregex (Levy & Andrew, 2006) expressions were used to find instances of each idiom. Matched literal phrases were extracted using Tregex to find sentences that included a phrase with the same syntactic structure as the idiom and the same verb. For example, to obtain matches for *bite the dust*, the procedure extracted sentences where the lemma *bite* had a direct object that was not *dust*.

The number of instances of the matched phrases ranged from several hundred to the tens of thousands (e.g., for verb object phrases beginning with *take* or *have*), with the majority falling in the range of a few hundred to a few thousand. There were fewer sentences containing idioms, with all 18 idioms having fewer than 100 instances.[3] For the non-idiomatic matched phrases, if there were more than 500 instances, only the first 500 sentences were included in the experiment.

[3]The corpus used in this experiment does not distinguish between idiomatic and literal uses of surface idioms; all instances are assumed to be idiomatic. Based on a manual check, it appears that very few instances of the surface idioms are literal uses, but this is nevertheless a drawback of the methodology.

5.3.2 Defining the sources and target

In order to compute information measure J, it is necessary to define the sources and target in the experimental setup. The definition of J is repeated below, where a distribution on meanings M is the target and the verb v and noun n are the sources.

$$J(M; v, n) = \sum_{m} p(m \mid v, n) \log \frac{p(m \mid v, n)}{p(m)} \tag{5.9}$$

We need to provide definitions for the probabilities $p(m \mid v, n)$ and $p(m)$, but in order to do so we must operationalize M. Since M is a random variable that ranges over the meanings of a particular phrase, we do not have direct access to it. In operationalizing M, I take as a starting point the idea that the meaning of a phrase largely determines which surrounding linguistic contexts the phrase can appear in.[4] For example, if the phrase is the idiom *spill the beans*, then the surroundings contexts might look something like the following:

My sister might ___ the ___ about the surprise party.

Don't ___ the ___ or you'll be sorry.[5]

...

The distribution over possible surrounding contexts for a particular interpretation of the phrase (i.e., idiomatic or literal) can be seen as a proxy for the distribution over meanings of that interpretation of the phrase. I further assume that the context words that are most directly determined by the meaning of the verb-object phrase are the content words rather than the function words, since function words serve primarily grammatical purposes, or else have meanings that are compatible with a vast range of sentences. By focusing solely on the content words, I hope to target the aspects of a sentence's meaning that are likely to change significantly based on the meaning of the verb-object phrase inserted into the sentence. To distinguish between the two types of context words, I use NLTK's list of `stopwords` (Bird et al., 2009). The stopwords, along with their positions in the sentence, can be thought of us containing whatever information is provided by the grammatical template of the sentence. We can formalize the above as follows.

[4] This idea is closely related to the Distributional Hypothesis in semantics, which says that similarity in linguistic distribution reflects similarity in meaning (Harris, 1954).

[5] As mentioned earlier in this chapter, I have left the determiner *the* in the sentence as part of the context, but it could instead be considered part of the idiom.

Let $\vec{w}$ be a vector containing the content words in the surrounding context of a particular occurrence of the verb-object phrase of interest. Each of the words in this vector is annotated with its position in the sentence. Let $\vec{s}$ be a vector containing the stopwords in the context, again where each element is annotated with its position. Then $p(\vec{w} \mid \vec{s}, m)$ is the probability of a particular set of content words in the surrounding context given the specific meaning m that the verb-object phrase carries in that particular context and the stopwords used in the sentence. The marginal probability of the surrounding content words $\vec{w}$ given the surrounding stopwords $\vec{s}$ is obtained by summing over the meaning variable M, which ranges over the possible intended meanings for an idiomatic or literal use of the phrase.

$$p(\vec{w} \mid \vec{s}) = \sum_{m} p(\vec{w} \mid m) p(m \mid \vec{s}) \tag{5.10}$$

Though we do not have direct access to the phrase's meaning, we can observe how the distribution $p(\vec{w} \mid \vec{s})$ changes when conditioned on particular verb-noun combinations that might be used to express the intended meaning. The posterior predictive distribution, then, is $p(\vec{w} \mid \vec{s}, v, n)$. Under this setup, the quantity $J(M; v, n)$ becomes, for a particular meaning:

$$J(\vec{W}; v, n) = \sum_{\vec{w}} \sum_{\vec{s}} p(\vec{w} \mid \vec{s}, v, n) \log \frac{p(\vec{w} \mid \vec{s}, v, n)}{p(\vec{w} \mid \vec{s})} \tag{5.11}$$

Computing $p(\vec{w} \mid \vec{s}, v, n)$ and $p(\vec{w} \mid \vec{s})$

To compute $p(\vec{w} \mid \vec{s}, v, n)$ and $p(\vec{w} \mid \vec{s})$, I use the masked language model XLNet (Yang et al., 2019), which allows one to obtain probabilities of words in context. Since XLNet is a masked language model, we can replace any number of words with [MASK] tokens and compute the probabilities of the remaining words. For example, in the sentence below, *spill* and *beans* have been replaced with mask tokens.

My sister might [MASK] the [MASK] about the surprise party.

I use the same masking procedure described in Chapter 3, which approximates marginalizing over the masked slot—i.e., of considering all possible insertions into that slots and averaging over them. To condition on a word, it suffices to leave the word in the sentence, unmasked. To compute

$p(\vec{w} \mid \vec{s}, v, n)$, I leave the verb, noun, and stopwords unmasked, and use the language model to predict the sequence of content words in the context. To compute $p(\vec{w} \mid \vec{s})$, I mask the verb and noun, leave the stopwords unmasked, and use the language model to predict the content words.

Estimating $J(\vec{W}; v, n)$ via sampling

This definition of $J(\vec{W}; v, n)$ is an expectation with respect to $\vec{W}$. To compute the true value of this expectation, one would have to loop over all possible content word sets and all possible stopword contexts, but this is not feasible. A natural approach to estimating this quantity would be to use Monte Carlo; in practice, I approximate Monte Carlo by using a stratified convenience sample consisting of the true contexts that occurred in the corpus for a particular verb-object phrase. Each context includes the sentence containing the verb-object phrase (but excluding the verb and noun), plus the preceding and following sentence in the corpus. For each of these contexts, $\vec{w}$ is the set of content words and $\vec{s}$ is the set of stopwords excluding the noun and verb. This sampling method rests on the assumption that the true contexts that occurred are highly probable in the distribution over possible contexts given a certain intended phrase meaning, accounting for a great deal of the probability mass.

5.4 Results and discussion

I first present results of the idiom and non-idiom corpora overall, followed by results on individual idioms within the idiom corpus, focusing on what we can conclude about both the methodology used and about the particular idioms in the corpus.

The first analysis compares the lower and upper bounds on the degree of compositionality score for the idiomatic and non-idiomatic corpora overall. Both of the bounds for the idiomatic corpus were lower than the corresponding bound for the non-idiomatic corpus. Figure 5.1 shows the compositionality range for the idiomatic and non-idiomatic corpora. Recall that the degree of compositionality measure is a proportion of the total information $J(v, n; M)$, so the true compositionality score should fall between 0 and 1. The lower bound, which is the proportion of $J(v, n; M)$ that the larger of $J(v; M)$ or $J(n; M)$ makes up, should always be between 0 and 1, whereas the upper bound, which double counts the redundancy, could in principle be as high as 2 (reaching its

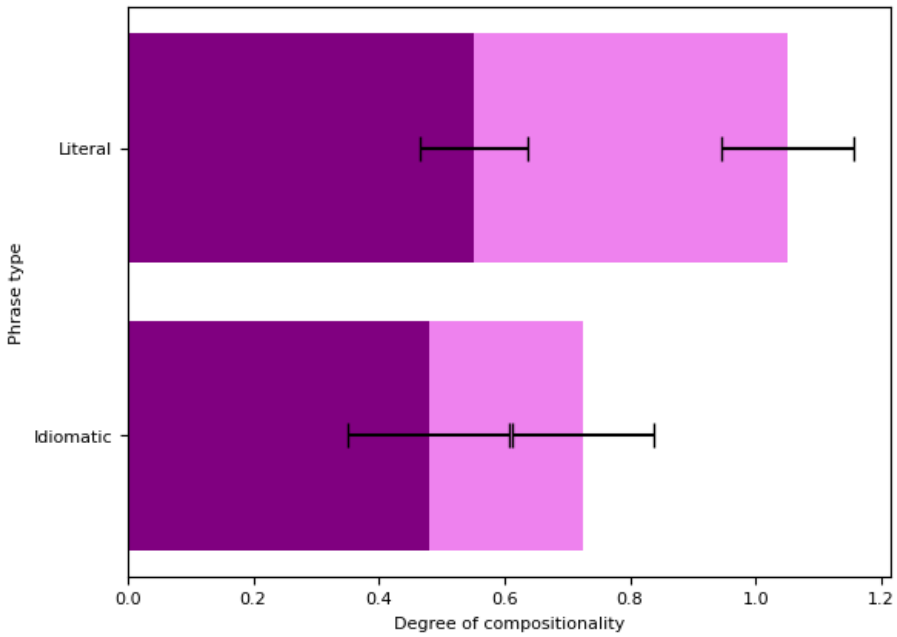

Figure 5.1: The compositionality score range (in light purple) between the upper and lower bounds for the idiomatic and non-idiomatic corpora.

maximum when the entirety of the information in the decomposition is redundant).[6]

In Figure 5.1 we see that the lower bound for the idiom corpus falls at 0.48, and the upper bound at 0.73. For the non-idiom corpus, the lower bound falls at 0.55 and the upper bound at 1.05. Ninety-five perfect confidence intervals for each bound were generated by repeatedly sampling ($n = 1000$) from the observations to estimate the standard error. A t-test comparing the lower bounds did not reveal a significant difference between means for the two groups ($p = 0.6$), but there was a nearly significant difference for the upper bounds ($p = 0.07$).

One thing to notice about these results is that the idioms, despite having more synergy overall than non-idioms, still have a large portion of their meaning made up of unique and/or redundant information. This is in line with one of the main motivating observations for this work, which is that the words in an idiom contribute partial compositional meaning to the idiomatic interpretations of the full phrase. Likewise, the non-idiomatic phrases involve some synergy, though not as much as the idioms do. Overall, the fact that the compositionality range is shifted closer to 1 for literal phrases than for idioms provides tentative evidence that the proposed compositionality measure

[6]Three of the phrases ended up with values for the bounds that were outside of what should be the possible ranges: *lead the field*, *close ranks*, and *deliver the goods*. These phrases have been excluded from the calculations.

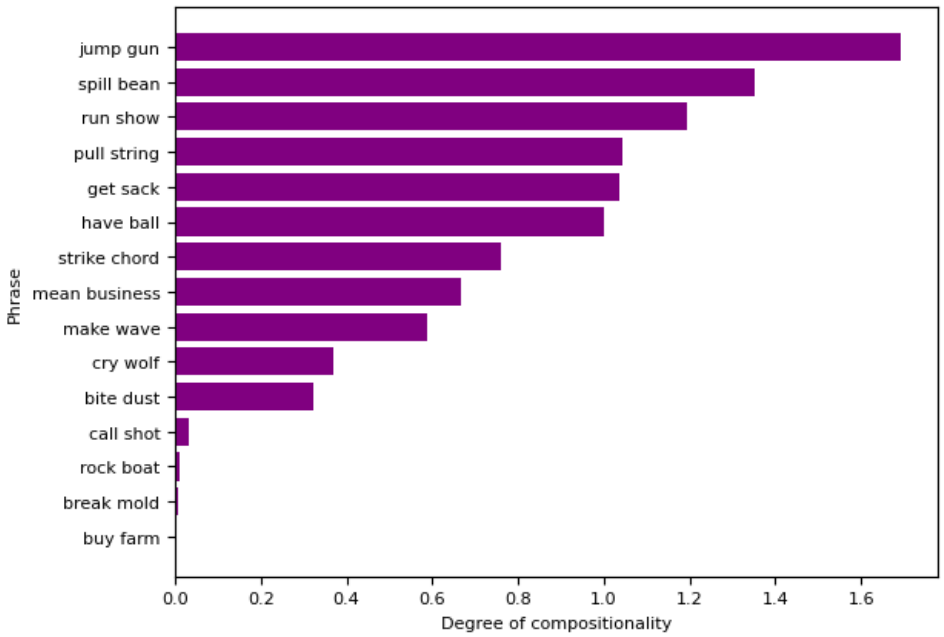

Figure 5.2: The mean of the upper and lower bounds on compositionality for individual idioms.

behaves as intended. At the same time, merely comparing the means for the two corpora may obscure more complicated patterns in the data; we also want to use our measure of compositionality to place individual phrases on a spectrum of how compositional they are. It is therefore useful to look at the results for individual idioms in our idiom corpus. Figure 5.2 displays the midpoint of the compositionality range for each idiom.

At the bottom of Figure 5.2 are the idioms that come out as the least compositional, while those at the top came out as the most compositional. Some of the placements of these idioms seem intuitive; for example, *buy the farm* has the idiomatic meaning "die," which has nothing obviously in common with the meanings of *buy* and *farm*, so it makes sense for its compositionality to be near zero. At the high-compositionality end of the spectrum we see *jump the gun*, meaning "act too early." This idiom has a highly compositional metaphorical interpretation—of springing to action before the gun has gone off in a race—that is clearly linked to the idiomatic meaning.

Overall informativeness in idioms versus non-idioms

This study has so far compared the relative size of the compositional portion of the total amount of information $J(\vec{W}; v, n)$ in idioms versus non-idioms, but we can also compare the total amounts

of information in the two groups as approximated by our procedure. In Chapter 4, we saw that the overall informativeness (as measured by surprisal, which is closely linked to J—see Section 2.6) of idiomatic phrases is lower than that of literal phrases. Here again we see a difference in total informativeness, this time measured as $J(\vec{W}; v, n)$, with literal phrases having more information overall than idiomatic phrases (1.38 bits to 1.04 bits, respectively). One possible reason for this is that many idioms seem, at least intuitively, to be longer ways of expressing concepts that can be paraphrased by a shorter utterance (e.g., *kick the bucket* = *die*, *shoot the breeze* = *chat*, etc.). While it is true that often the single-word paraphrase does not quite capture all of the nuances of the idiom's meaning, it may be that these idioms convey less information than an average literal verb-object phrase, in which each of the component words plays a significant and largely independent part in narrowing the space of possible meanings.

5.5 Conclusion

As emphasized at the beginning of this chapter, the study presented here is a pilot, and many simplifying assumptions have been made. The major limitations are listed below. However, even with these limitations, the fact that a conservative approximation of the proposed compositionality measure shows a clear difference between idioms and non-idioms provides suggestive evidence in favor of the PID-based measure, which offers a way of quantifying the compositional contribution of individual words/expressions.

Limitations of the current study

- The measure was tested on a small number of instances, particularly among the set of idioms.
- The scores produced by the implementation were not compared against human judgment data.
- The masking procedure served only as an approximation of marginalization.
- A simplified sampling procedure was used to approximate Monte Carlo.
- The scores are subject to any peculiarities in the language model probability estimates and/or the model's internal representations of word meanings.

Future work will ideally address these limitations, paving the way for further studies on the topic. For example, while the present study compares idioms to syntactically matched literal phrases, another way of validating the measure would be to compare idiomatic versus literal uses of the same phrase to see if the measure can properly distinguish them. Additionally, while this study focused on idioms as a representative sample of low-compositionality phrases, the method could be extended to study the relative amounts of compositionality in other constructions that have been at the heart of debates about the compositionality of natural language.

Preface to Chapter 6

The PID-based approach to measuring compositionality does not impose any requirements on the size of the linguistic units it operates on. Chapter 5 tested the approach at the phrase level, looking at the contributions of words to idiomatic and non-idiomatic multiword phrases. Chapter 6 zooms in on the sub-word domain and measures systematicity in the relationship between individual morphemes and the meanings they contribute. Differences in morphological systematicity have been described in the linguistics literature as a divide between agglutinative (highly systematic) and fusional (less systematic) languages. PID offers an opportunity to place morphological systems along a spectrum of systematicity.

The study reported in this chapter was designed and conducted before the development of the compositionality measure proposed in Chapter 2, and there are a few key differences in the way the problem is set up. First, the proportion of synergistic information is taken to be a measure of the amount of fusion in a morphological system (whereas previously in this thesis, its inverse—the proportion of non-synergistic information—was taken as a measure of compositionality). Second and more significantly, the assignment of source and target variables is done differently. In Chapters 2 and 5, I described the goal as that of decomposing the amount of information that individual linguistic forms contribute about the meaning of a larger expression; therefore we defined linguistic forms as sources and meanings as targets. Turning to morphology, the analogous approach would be to treat individual morphemes as sources and the meaning of the complete word as the target. However, there is no obvious way to decompose highly fusional words into separable morphemes—this is, of course, what it means to be fusional. With inflectional morphology, it is the meaning side that is more straightforwardly decomposable. For example, both a highly fusional and a highly agglutinative language can express a noun with the decomposable meaning *dog + plural*

+ *nominative*, and it is the distribution of these units of meaning across the linguistic form that determines the level of fusionality. For this reason, the study in Chapter 6 takes units of meaning as source variables and character slots within the linguistic form as target variables, with the decomposition computed separately for each target slot.

A further difference in Chapter 6's approach is that mutual information rather than J is used as the quantity to decompose. Instead of the sources being outcomes, they are random variables corresponding to semantic categories (e.g., CASE and NUMBER), each of which can take on a (small) set of values. Each target slot takes on the value of whichever linguistic form occurs in that slot. Finally, a definition of unique information from Bertschinger et al. (2014) is used to compute the decomposition. This definition is closely related to Kolchinsky (2022)'s definition of union information; see Kolchinsky (2022) for further discussion of their properties.

Chapter 6

Measuring morphological fusion using Partial Information Decomposition

6.1 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.

In Table 6.1, 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

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 6.1: 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.

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 P. Williams & 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.

6.2 Partial information decomposition

6.2.1 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:

$$I(M;F) = \sum_{m \in M} \sum_{f \in F} p(m,f) \log \frac{p(m,f)}{p(m)p(f)}. \tag{6.1}$$

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

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

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

Table 6.2: (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.

Table 6.2. 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 information based on how each meaning variable contributes information—on its own, redundantly with other variables, or jointly with other variables.

6.2.2 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. P. Williams & Beer (2010)'s PID framework provides an influential solution to the decomposition problem; we briefly summarize the framework here.

P. Williams & 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

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

particular quantity of information, summarized in Table 6.3.[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.

P. Williams & Beer show that the collections can be naturally structured into a partially-ordered lattice, shown for the case of two source variables in Figure 6.1. 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 6.2. 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 6.3 (and the same for S_2, in Equation 6.4).

$$I(S_1, S_2; T) = R_{1,2} + U_1 + U_2 + S_{1,2} \tag{6.2}$$

$$I(S_1; T) = R_{1,2} + U_1 \tag{6.3}$$

$$I(S_2; T) = R_{1,2} + U_2 \tag{6.4}$$

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

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

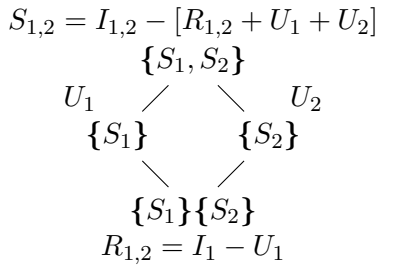

Figure 6.1: Partial information lattice for the case of two source variables. The equations at each node are abbreviated versions of equations (6.2)–(6.4), showing how to solve for redundant, unique, and synergistic information, starting at the bottom of the tree.

solve.

Gutknecht et al. (2021), building on P. Williams & Beer (2010), show that with a definition of either redundant information or unique information, it is possible to solve the system of equations 6.2–6.4 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., P. Williams & Beer, 2010; Bertschinger et al., 2014; Finn & 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).

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:

$$U_1 = \min_{Q \in \Delta_P} I_Q(S_1; T \mid S_2) \tag{6.5}$$

where

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

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 6.2, we get a unique information of 1 for the fully systematic example and 0 for the less systematic example. In Section 6.3.2 we define a measure of morphological fusion based on the Bertschinger et al. (2014) formulation.

6.3 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 6.4. All of the languages in our experiment have noun paradigms with exactly two non-stem meaning feature categories: CASE and NUMBER.

6.3.1 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.

Source Meaning Variables Consider the morphological paradigm for the Russian noun КОТ in Table 6.4, 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)

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 6.4: Subset of paradigm for the Russian noun кот.

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 & 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 6.5.

Our application differs from the original PID formulation in that we are dealing with multiple target variables. In Section 6.3.2 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 F_1 represent the stem's meaning (e.g., *cat*) and the stem's form (e.g., кот), respectively (Table 6.5).

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 6.5: Random variable structure for three word forms in Russian.

6.3.2 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 F, 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:

$$\bar{U}_{1,2\to F} \coloneqq \frac{U_{1\to F} + U_{2\to F}}{I(M_1, M_2; F)} \tag{6.6}$$

$$\bar{R}_{1,2\to F} \coloneqq \frac{R_{1,2\to F}}{I(M_1, M_2; F)} \tag{6.7}$$

$$\bar{S}_{1,2\to F} \coloneqq \frac{S_{1,2\to F}}{I(M_1, M_2; F)} \tag{6.8}$$

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:

$$\bar{h} = \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \frac{1}{|\mathcal{F}|} \sum_{F \in \mathcal{F} - \{F_1\}} h_{1,2\to F} \tag{6.9}$$

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

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.

6.4 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.

6.4.1 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.

6.4.2 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 6.2 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%.

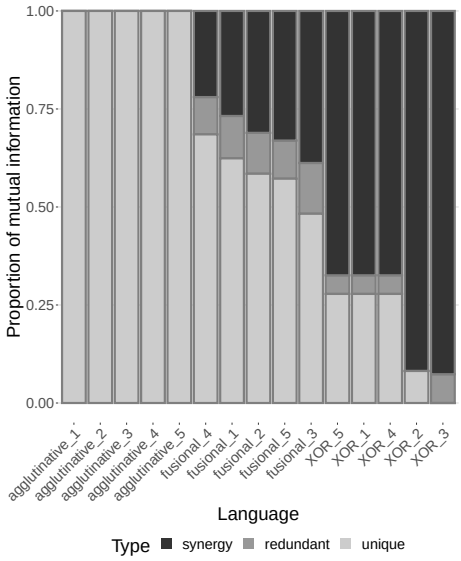

Figure 6.2: Results of partial information decomposition on noun paradigms in baseline artificial languages. The languages are sorted by relative amount of synergy.

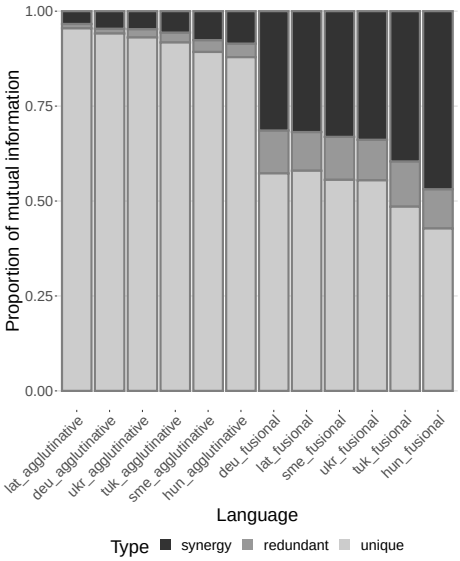

Figure 6.3: Results of partial information decomposition on noun paradigms in linguistically-controlled artificial languages. The languages are sorted by relative amount of synergy.

6.4.3 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 6.3. 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.

6.4.4 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 6.4, 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 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 H.1 in Appendix H), and (2) a plot of the average number of suffix slots in each language (Figure H.2 in Appendix H). 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.

6.5 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,

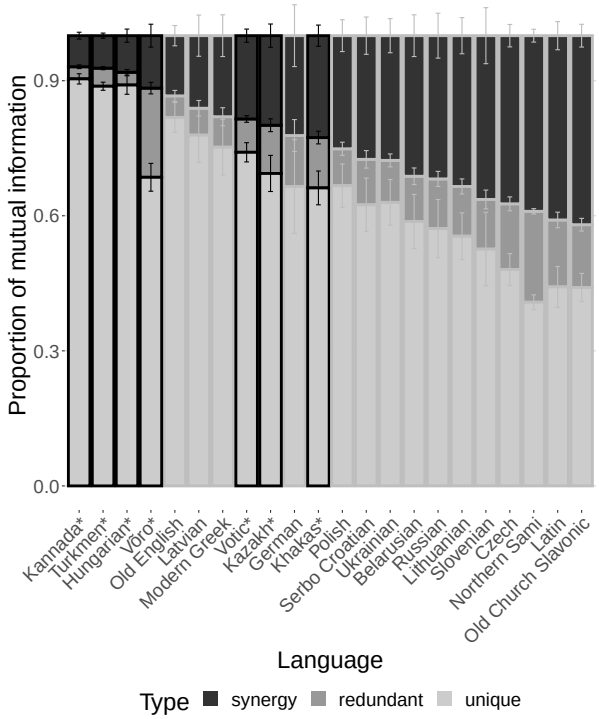

Figure 6.4: Results of partial information decomposition on noun paradigms in 22 languages. The languages are sorted by relative amount of synergy. Asterisks and dark borders represent languages labeled as agglutinative in UniMorph.

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 6.6. 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 6.7. 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.

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 6.6: 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 6.7: PID values (unique, redundant, synergistic) for the Latin paradigm in Table 6.6, unnormalized.

6.6 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 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 (A. 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:

$$\phi(w) = -\log p(w \mid \mathcal{L}_{-\sigma}, \sigma, \ell) \tag{6.10}$$

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.

6.7 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.

Chapter 7

Conclusion

I conclude with a summary of the main contributions of this thesis, followed by an overview of possible avenues for future work.

7.1 Summary of thesis

This thesis has taken up the issue of compositionality in language. For decades, compositionality has been a subject of debate in linguistics and related fields, with arguments centering around two main questions, repeated here:

(40) What does compositionality refer to?

(41) Is language compositional? (To what extent?)

Chapter 2 reviewed the literature on formalizing compositionality and proposed a new formalization based on insights from information theory. I defined a measure of the *degree of compositionality* of an utterance, which can be applied at any level of linguistic granularity. The framework and definition that I have proposed provide an answer to the first question above, allowing us to set about answering the second.

Chapter 3 showed that idioms can be characterized as occurring at the intersection of two cognitive processes: one that allows for a word to have an unconventional meaning, where the amount of unconventionality exists on a spectrum, and one that allows for the storage and reuse of particular phrases. Chapter 4 looked at idioms from a psycholinguistic perspective, showing (in line

with previous literature) that idioms are spoken more quickly than literal phrases, and that this can be primarily explained by surprisal rather than the conventionality of the component words' meanings. This finding provides evidence against the idea that compositionality is directly reflected in processing time.

Chapter 5 presented a pilot study evaluating the proposed compositionality measure on a corpus of idioms and matched non-idiomatic phrases, with results in the expected direction, and Chapter 6 showed that a modified version of the measure succeeds at capturing distinctions in systematicity in morphology. The success of the measure in both word- and phrase-level domains provides evidence for its ability to measure compositionality in language overall (and potentially also in domains outside of language).

7.2 Future directions

Using the definition of compositionality proposed in this thesis, it is possible to investigate how compositionality impacts grammatical structure and vice versa. If we can characterize the grammatical configurations where non-compositionality tends to arise, then we can make further progress in understanding (1) how form-meaning mappings are represented in the mind, and (2) how non-compositional utterances are learned and processed. Below are a few types of studies that could be carried out to achieve this aim.

7.2.1 Syntactic flexibility of idioms

Idioms vary when it comes to the syntactic configurations in which they can appear. Nunberg et al. (1994) has suggested that there are two distinct classes of idioms—those that are syntactically flexible and those that are not—with *Spill the beans* being an example of the former class ("The beans were spilled," "She spilled a few beans"), and *kick the bucket* being an example of the latter. Nunberg et al. (1994) hypothesized that idioms whose individual words can be understood in a metaphorical relation to the meaning of the idiom are the more flexible kind (e.g., in *spill the beans*, *spill* = REVEAL and *beans* = INFORMATION). More recently, Stone (2016) has shown that the empirical phenomenon is actually more complex: there is an implicational hierarchy of syntactic constructions, such that if an idiom can appear in a particular construction, then it can also appear

in every type of construction beneath that one on the hierarchy.

The PID-based definition of compositionality proposed in this thesis could be used to investigate the relationship between the syntactic flexibility of idioms and their degree of compositionality. For example, compositionality scores could be computed for the idioms in our validation study from Chapter 3, with stimuli designed such that each idiom appears in various syntactic configurations (e.g., passivization, pluralization, adjectival modification, relative clause formation). Participants could provide judgments about whether the idiomatic meaning is available in a given configuration.

The results of such a study would help characterize the relationship between an idiom's flexibility and its degree of compositionality. If Nunberg et al. (1994) is correct, then we should see the most inflexible idioms having the lowest proportion of synergy. Since the compositionality measure places phrases along a spectrum, it is well suited for studying the effect of compositionality on syntactic flexibility, as syntactic flexibility is also said to be a matter of degree. If it turns out that phrases with a high proportion of synergy resist syntactic transformation, then this may suggest that they are stored and accessed in the mind as a unit, or perhaps that syntactic transformations apply most readily to words that are associated with independently-contributing (i.e., non-synergistic) meanings.

7.2.2 Syntactic correlates of compositionality

One of the main advantages of the proposed measure of compositionality is that it allows us to quantify the degree of compositionality of any word, phrase, or larger utterance, as long as the component parts are specified. This opens up the opportunity of probing a corpus of text for places within syntactic structure where the levels of non-compositionality tend to be particularly high or particularly low, and then generate new predictions based on those findings. Doing so would allow us to build a corpus annotated for compositionality scores at a fine-grained level, making it a useful resource for testing new predictions. For example, one could use the Tregex tool to automatically annotate each sentence in an existing corpus with its syntactic tree structure (Levy & Andrew, 2006), and then, for each sentence, compute the degree of compositionality score at each node in the tree structure.

With such a corpus, we could ask a variety of questions, such as whether verb phrases tend to be highly synergistic, which might help explain why verb-object idioms are so common. It is frequently

reported that there are no subject-verb idioms (e.g., O'Grady, 1998); one could investigate whether the node linking subjects with their predicates tends to have a lower proportion of synergy. If it turned out that different amounts of synergy are associated with different parts of a sentence, this could provide evidence for or against particular proposals in syntactic theory. As an example, there is debate in the literature about the existence of *phases*, which are parts of the structure that, it is argued, cannot be accessed once they are built (Chomsky, 2000). If it turns out that synergistic meaning contributions tend not to cross the posited phase boundaries, then this could be suggestive evidence that the boundaries exist.

It would also be interesting to observe whether there is a sharp distinction between structures that tolerate high levels of synergy and those that do not, or if syntactic structures, like individual phrases, are spread along a spectrum of compositionality. Additionally, Chapter 3 describes a finding that the head words of idioms tend to have more conventional meanings than their dependents; we could use the PID-based definition of compositionality to further investigate this phenomenon.

7.2.3 Correlation or trade-off between compositionality in morphological domains with a language?

The study in Chapter 6 showed that the relative amount of synergistic information in a morphological noun system can serve as a measure of the fusionality of the system. It would be interesting to see whether languages that exhibit a high degree of fusionality with nouns exhibit the same in other areas, such as verbs and adjectives. If so, this could suggest that languages with a consistent amount of systematicity/compositionality across different domains are easiest to learn. On the other hand, we might expect to see a trade-off between the amounts of systematicity of different areas, such that different languages end up having, on average, similar overall amounts.

7.2.4 Change over time

We could gain insight into the question of how compositionality in language changes over time by comparing historical and modern corpora in a variety of languages. It would be particularly interesting to investigate the development of idioms over time in order to discern the factors that affect the lifespan of synergistic combinations. One hypothesis about language evolution is that language evolves for efficiency (e.g., Kanwal et al., 2017); however, it is not clear that this applies

in the case of idioms. For example, saying "leave no stone unturned" is less efficient than saying its literal meaning ("be thorough"). The results of an investigation into idiomatic use over time could shed light on, for instance, whether the words in an idiom tend to extend their meanings until the idiomatic use of a particular word becomes available in more contexts, or whether the words in an idiom become increasingly bleached of their literal meanings as the origin of the idiom disappears from cultural knowledge. It could of course be the case that different idioms show different trajectories.

Appendices

Appendix A

Target phrases

Target phrase	Type	Target phrase	Type
deliver the goods	VO	swimming pool	NN
run the show	VO	cash cow	NN
rock the boat	VO	foot soldier	NN
call the shots	VO	attorney general	NN
talk turkey	VO	hit list	NN
cut corners	VO	soup kitchen	NN
jump the gun	VO	bull market	NN
have a ball	VO	boot camp	NN
foot the bill	VO	message board	NN
break the mold	VO	gold mine	NN
pull strings	VO	report card	NN
mean business	VO	comfort food	NN
raise hell	VO	pork barrel	NN
close ranks	VO	flower girl	NN
strike a chord	VO	hit man	NN
cry wolf	VO	blood money	NN
lose ground	VO	cottage industry	NN
make waves	VO	board game	NN
clear the air	VO	death wish	NN
pay the piper	VO	word salad	NN
spill the beans	VO	altar boy	NN
bite the dust	VO	bench warrant	NN
saw logs	VO	time travel	NN
lead the field	VO	love language	NN
take the powder	VO	night owl	NN
buy the farm	VO	life blood	NN
turn tail	VO	road rage	NN
get the sack	VO	light house	NN
hit the sack	VO	bid price	NN
kick the bucket	VO	carrot cake	NN
shoot the bull	VO	command line	NN
		stag night	NN
		husband material	NN

Target phrase	Type	Target phrase	Type
cold feet	AN	by and large	B
green light	AN	more or less	B
red tape	AN	bits and pieces	B
black box	AN	up and down	B
blue sky	AN	rise and fall	B
bright future	AN	sooner or later	B
sour grape	AN	rough and ready	B
green room	AN	far and wide	B
easy money	AN	give and take	B
last minute	AN	time and effort	B
hard heart	AN	pro and con	B
hot dog	AN	sick and tired	B
raw talent	AN	back and forth	B
hard labor	AN	day and night	B
broken home	AN	wear and tear	B
fat chance	AN	nut and bolt	B
dirty joke	AN	tooth and nail	B
happy hour	AN	on and off	B
high time	AN	win or lose	B
rich history	AN	food and shelter	B
clean slate	AN	odds and ends	B
stiff competition	AN	in and out	B
maiden voyage	AN	sticks and stones	B
cold shoulder	AN	make or break	B
clean energy	AN	part and parcel	B
hard sell	AN	loud and clear	B
back pay	AN	cops and robbers	B
deep pockets	AN	short and sweet	B
broken promise	AN	safe and sound	B
dead silence	AN	black and blue	B
blind faith	AN	toss and turn	B
tight schedule	AN	fair and square	B
brutal honesty	AN	heads or tails	B
bright idea	AN	hearts and flowers	B
kind soul	AN	rest and relaxation	B
bruised ego	AN	flesh and bone	B
		life and limb	B
		checks and balances	B
		fast and loose	B
		high and dry	B
		pots and pans	B
		now or never	B
		hugs and kisses	B
		bread and butter	B
		risk and reward	B
		cloak and dagger	B
		nickel and dime	B
		rhyme or reason	B
		leaps and bounds	B
		live and learn	B
		peace and quiet	B
		song and dance	B
		pins and needles	B
		sugar and spice	B
		neat and tidy	B
		step by step	B
		lost and found	B
		old and grey	B

Appendix B

Literalness rating model results

Table B.1: Model results table with human literalness rating as the dependent variable, using `lmer`

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
Intercept	0.051	0.019	1.655	0.049
Conv	0.185	0.050	3.725	< 0.001
Head(False)	0.015	0.014	1.050	0.147
Conv:Head(False)	0.073	0.053	1.376	0.084

$n = 4945$

Table B.2: Model results table for model described in Section 3.5.2, with contingency score as the dependent variable, using `lmer`

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
Intercept	4.949	0.114	43.379	< 0.001
Target(True)	1.253	0.165	7.587	< 0.001
Class(VO)	-0.195	0.200	-0.975	0.165
Class(AN)	-0.662	0.201	-3.297	< 0.001
Class(B)	1.796	0.179	10.045	< 0.001
Target(True): Class(VO)	0.501	0.303	1.654	0.049
Target(True): Class(AN)	-0.896	0.286	-3.135	< 0.001
Target(True): Class(B)	1.394	0.247	5.641	< 0.001

$n = 99573$

Table B.3: Model results table for model described in Section 3.6, with conventionality score as the dependent variable

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
Intercept	0.163	0.035	4.614	< 0.001
PhraseConv	0.526	0.036	14.453	< 0.001
Class(VO)	0.196	0.065	3.020	0.003
Class(AN)	-0.135	0.063	-2.153	0.032
Class(B)	-0.010	0.064	-0.150	0.881
Head(False)	-0.326	0.050	-6.525	< 0.001
PhraseConv:Class(VO)	-0.250	0.062	-4.043	< 0.001
PhraseConv:Class(AN)	0.117	0.069	1.683	0.093
PhraseConv:Class(B)	0.116	0.068	1.694	0.091
PhraseConv:Head(False)	0.476	0.051	9.247	< 0.001
Class(VO):Head(False)	-0.392	0.092	-4.271	< 0.001
Class(AN):Head(False)	0.271	0.089	3.044	0.002
Class(B):Head(False)	0.019	0.091	0.212	0.832
PhraseConv:Class(VO): Head(False)	0.500	0.087	5.717	< 0.001
PhraseConv:Class(AN): Head(False)	-0.233	0.098	-2.380	0.018
PhraseConv:Class(B): Head(False)	-0.232	0.097	-2.396	0.017

$n = 584$

Appendix C

Target phrase visual analysis

Below is a list of the target phrases that landed in each of the quadrants in Figure 3.3, for those phrases that occurred at least 30 times in the corpus.

Top left	**Top right**
black and blue	back and forth
black box	bits and pieces
bread and butter	boot camp
by and large	bright future
call the shots	deep pockets
checks and balances	deliver the goods
clear the air	far and wide
cottage industry	food and shelter
cut corners	heads or tails
day and night	high and dry
foot soldier	more or less
give and take	on and off
gold mine	part and parcel
happy hour	pull strings
have a ball	rise and fall
high time	rock the boat
in and out	run the show
loud and clear	song and dance
make or break	swimming pool
nuts and bolts	up and down
peace and quiet	
red tape	
safe and sound	
sick and tired	
soup kitchen	
sour grapes	
win or lose	

Bottom left	Bottom right
cold feet	blue sky
green light	board game
hard sell	bright idea
hit man	get the sack
hot dog	green room
last minute	hit list
lose ground	report card
mean business	time and effort

Appendix D

Replication on a separate dataset

To confirm that the results in Chapter 3 are not simply an artifact of the dataset we used, we replicated the study on a second dataset, which is the set of phrases used in the idiom detection work of Fazly et al. (2009). We did not have any hand in choosing the phrases in this dataset, and it has very little overlap with our own. We once again fail to find evidence that the two dimensions of conventionality and contingency are correlated with one another in this set of phrases ($r(24)$ = -0.276, p = 0.172), and we see a similar spread of data across the four quadrants, shown in Figure D.1.

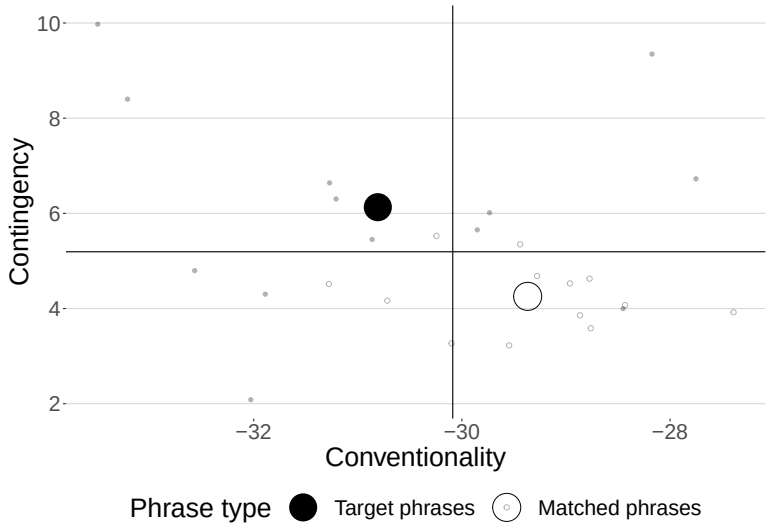

Figure D.1: Contingency and conventionality values of target and matched phrases. Large circles are average values of all target (black) and matched (white) phrases.

Appendix E

Union information

As discussed in Chapter 2, a definition of unique, redundant, or union information is necessary for performing partial information decomposition and thus for computing the true compositionality score of an expression, rather than just bounds on the score. I propose a method for approximating union information (i.e., the sum of unique and redundant information), inspired by a definition given in Kolchinsky (2022) that has been shown to behave intuitively in a wide range of cases.[1] Union information is defined by Kolchinsky (2022) with respect to mutual information, whereas in the present study I replace mutual information with information measure J. Regardless of which information measure is chosen, union information gives the sum of the unique and redundant contributions, i.e., the numerator of the proposed compositionality measure: $U_v + U_n + R$. The definition of union information as given in Kolchinsky (2022) (with mutual information) is the least amount of information obtainable about the target from a random variable Q that is more informative than each of the sources. The definition is stated as follows:[2]

$$I_{\cup}(X_1, ..., X_n; Y) \coloneqq \inf_Q I(Q; Y) \text{ such that } \forall i X_i \sqsubset Q \tag{E.1}$$

where there is some ordering under such that $X_i \sqsubset Q$ indicates that X_i is less informative about Y than Q is. In other words, the more informative variable contains all of the information about

[1] Kolchinsky (2022)'s definition of union information generalizes a commonly-used measure proposed by Bertschinger et al. (2014), which I use in Chapter 6 to measure morphological systematicity.

[2] Finding a single variable Q that minimizes the mutual information $I(Q; Y)$ is not guaranteed, which is why the definition specifies infimum instead of minimum.

Y that is contained in the less informative variable. The quantity $I(Q;Y)$ in the definition is the mutual information between Y and the smallest random variable that is more informative about Y than all of the Xs are. This definition is based on the set-theoretic concept of the union of random variables, and it can be useful to understand it in this way. For the case of two random variables, the union information represents the set of information contained in one or the other or both variables individually, and excluding any information that requires simultaneous knowledge of both. See Kolchinsky (2022) for a detailed explanation of the analogy between PID and set theory.

To implement the definition of union information, it is necessary to select an ordering relation $\sqsubset$ on a set of random variables relative to the target; the choice of ordering may "depend on the operational setting and scientific domain in which the PID is applied" (Kolchinsky, 2022, p. 5). One possibility, proposed by Kolchinsky (2022), is known as the Blackwell order (Blackwell, 1953), and this serves as inspiration for the experimental setup used here. Given random variables B, C, and target Y, the Blackwell ordering relation $B \prec_Y C$ is defined as:

$$B \prec_Y C \text{ iff } P_{B|Y}(b \mid y) = \sum_c \kappa_{B|C}(b \mid c) P_{C|Y}(c \mid y) \text{ for some channel } \kappa_{B|C} \text{ and all } b, y. \tag{E.2}$$

This ordering relation is relative to random variable Y. The definition says that it is possible for the conditional distribution $P(B \mid Y)$ to be generated by applying a channel $\kappa(B \mid C)$ to a sample from conditional distribution $P(C \mid Y)$. This channel is an arbitrary conditional distribution of C given Y that may be different from $P(C \mid Y)$. By the data processing inequality (see Cover & Thomas, 2006), $B \prec_Y C$ implies that B has less mutual information about Y than C does.

$$B \prec_Y C \implies I(B;Y) \leq I(C \mid Y). \tag{E.3}$$

In adapting union information to our setting, we first replace mutual information I with semi-pointwise information measure J between a particular verb v and noun n and the meaning M of the verb-noun phrase. As described in in Chapter 5, M is operationalized in such a way that we actually compute $J(\vec{w} \mid \vec{s}, m, n, v)$. A definition of union information can be used to compute the

compositionality score as follows:

$$\text{Compositionality}(\vec{W}; n, v) = \frac{\text{UnionInfo}(\vec{W}; n, v)}{J(\vec{W}; n, v)}. \tag{E.4}$$

By analogy with the definition of union information above, we specify an ordering for our setting. Instead of an ordering over random variables, we take an ordering over the set of strings that can be inserted in place of the verb-object phrase. Furthermore, we want to consider only those strings that are more informative about the meaning of the phrase (as approximated by the context) than either v or n alone. Consider a set of linguistic forms (e.g., words or phrases), each of which can be transformed into v and, separately, into n, by adding some amount of noise (i.e., sending the form through some channel). We want to find a form q for which we obtain an infimum of $J(\vec{W}; q)$.

$$\text{UnionInfo}(\vec{W}; q) \coloneqq \inf_q J(\vec{W}; q) \text{ such that } v \prec_{\vec{W}} q \text{ and } v \prec_{\vec{W}} q. \tag{E.5}$$

I propose that computing the information J between the context and a paraphrase of the verb-object phrase's literal meaning serves as an approximation of the infimum. Consider the case where the phrase of interest is an idiom such as *spill the beans*. In order to be informed about the meaning of this phrase, it is necessary to know that both of the words *spill* and *beans* have occurred. In choosing a substituting phrase that gets us close to the infimum, we want a phrase that is less informative about the meaning "reveal the secret" than the phrase *spill the beans* is, but more informative than either *spill* or *beans* on its own. In particular, we want the least informative such phrase. By choosing a phrase that employs (near-)synonyms for the literal meanings of the verb and the noun (such as *dump the legumes* or *leak the grains* for *spill the beans*), we may approach the infimum, on the assumption that each word in the new phrase is roughly as informative as its original counterpart considered in isolation.[3] The information $J(\vec{W}; q)$ using the literal paraphrase then becomes an approximation of the union information (though likely overestimates it). In the

[3]In a fully compositional phrase, it is the original phrase itself that acts as a minimizer, but we cannot simply consider the original phrase in the case of idioms, since it would be interpreted in context with the idiomatic meaning. We opt for a paraphrase of the literal meaning as the next best option.

case of an idiom, the literal paraphrase is predicted to be significantly less informative than the true phrase, yielding low union information. In the case of a non-idiomatic phrase, a literal paraphrase should be nearly as informative as the original phrase, yielding high union information. A version of the pilot experiment in Chapter 5 was conducted using this implementation of union information.

Results

Figure E.1 shows the means of the compositional information estimated for the idiomatic and non-idiomatic corpora. As expected, we see the idioms showing significantly less compositionality than the literal phrases. However, the mean value for the non-idiomatic phrases falls at 6.61, greatly exceeding the expected maximum of 1 and indicating that the replacement phrases used to estimate union information conveyed more information than the original literal phrases. The method of estimating union information is fairly imprecise, as it relies on choosing a single near-synonymous paraphrase of the original phrase's literal meaning, and it seems that the literal paraphrases, in failing to be entirely synonymous, caused a larger-than-expected change in belief about the meaning of the sentence. As for the idiomatic phrases, the mean score for these phrases falls at 0.45, which is slightly below the lower bound computed in the previous analysis but within the error bars. The fact that we see a clear difference between idioms and non-idioms despite using such a coarse estimation is once again promising for the PID-based approach, but more sophisticated estimation methods will likely yield more reliable results.

Figure E.2 shows the scores generated by this method for the individual idioms in our dataset. When evaluating these results, it is important to note the difficulty of selecting literal paraphrases for certain idiomatic phrases. Some of the idioms in our dataset are in fact not very idiomatic upon close inspection. For instance, in *get the sack*, the word *sack* has the meaning of a firing or dismissal, which is a possible meaning of the word outside the context of this particular phrase. In choosing a synonym for *sack*, I opted for a word that matched this sense—*dismissal*—but it is difficult to know how the language model represents the meanings of such words in isolation. When a particular sense of a word seemed obvious in the idiom, I tried to choose a synonym for that particular meaning of the phrase, but there were also idioms for which it was not entirely clear which literal meanings of the component words were being referenced.

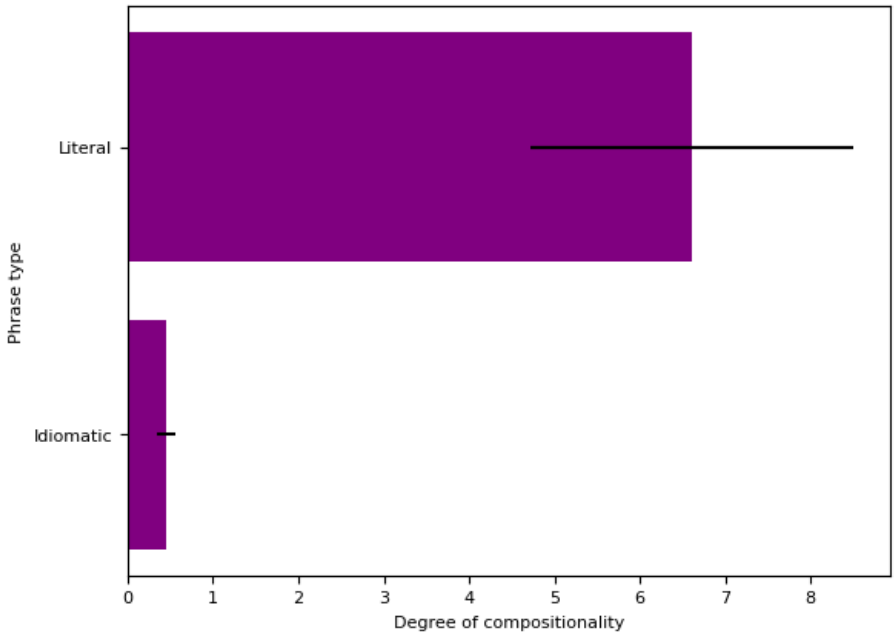

Figure E.1: The compositionality score estimated using union information for the idiomatic and non-idiomatic corpora.

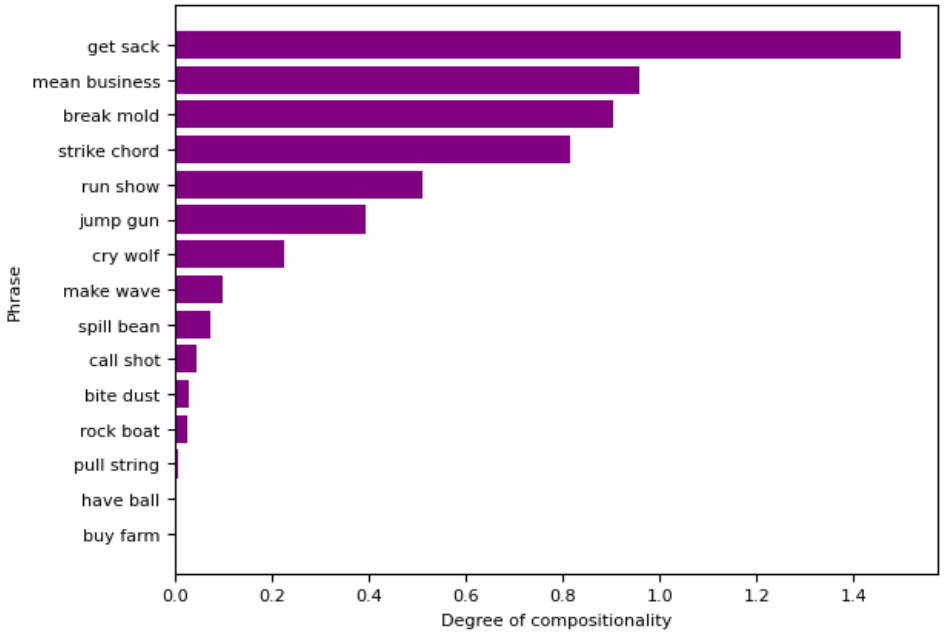

Figure E.2: The compositionality score estimated using union information for individual idioms.

Appendix F

Toy paradigm analysis

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 F.1: Toy language, noun 1.

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

Table F.2: 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.

Appendix G

Pseudocode for PID morphology study

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

Appendix H

Mutual information and suffix length analyses

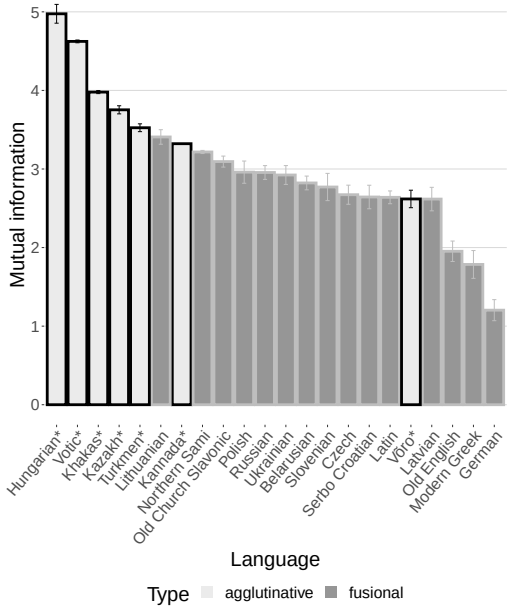

Figure H.1: Average amount of mutual information between meaning and form in the nominal paradigms of 22 languages. Asterisks and dark borders represent languages labeled as agglutinative in UniMorph.

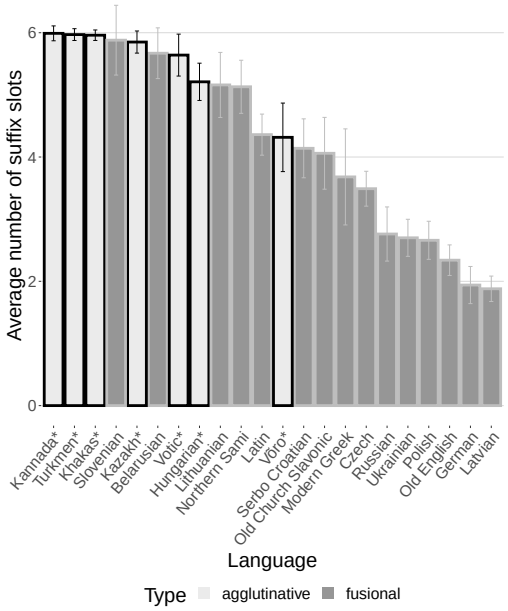

Figure H.2: Average suffix length in the nominal paradigms of 22 languages. Asterisks and dark borders represent languages labeled as agglutinative in UniMorph.

References

Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second Language Research*, *19*, 329–358.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, *12*, 5–32.

Bélanger, N., Baum, S. R., & Titone, D. (2009). Use of prosodic cues in the production of idiomatic and literal sentences by individuals with right- and left-hemisphere damage. *Brain & Language*, *110(1)*, 38–42.

Bertschinger, N., Rauh, J., Olbrich, E., & Jost, J. (2013). Shared information–new insights and problems in decomposing information in complex systems. In *Proceedings of the European Conference on Complex Systems 2012* (pp. 251–269). Springer.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., & Ay, N. (2014). Quantifying unique information. *Entropy*, *16*, 2161–2183.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Blachman, N. (1968). The amount of information that y gives about X. *IEEE Transactions on Information Theory*, *14*(1), 27–31.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, *24*(2), 265–272.

Bloomfield, L. (1933). *Language*. Henry Holt, New York.

BNC Consortium. (2007). *The British National Corpus, XML Edition*. Oxford Text Archive.

Bobrow, S., & Bell, S. (1973). On catching on to idiomatic expressions. *Memory & Cognition*, *1*, 343–346.

Bruening, B. (2020). Idioms, collocations, and structure: Syntactic constraints on conventionalized expressions. *Natural Language and Linguistic Theory*, *38*, 365–424.

Bruening, B., Dinh, X., & Kim, L. (2018). Selection, idioms, and the structure of nominal phrases without classifiers. *Glossa*, *42*, 1–46.

Brugman, C. (1988). *The story of over: Polysemy, semantics, and the structure of the lexicon.* Garland Press.

Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In *Lexical Ambiguity Resolution* (pp. 477–508). Elsevier.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, *27*, 668–683.

Caillies, S., & Butcher, K. (2007). Processing of idiomatic expressions: Evidence for a new hybrid view. *Metaphor and Symbol*, *22*, 79–108.

Cappelen, H., Lepore, E., & McKeever, M. (2005). Quotation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University.

Carnap, R. (1956). *Meaning and necessity.* Chicago: The University of Chicago Press.

Carrol, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, *63*, 95–122.

Carrol, G., Littlemore, J., & Dowens, M. G. (2018). Of false friends and familiar foes: Comparing native and non-native understanding of figurative phrases. *Lingua*, *204*, 21–44.

Carston, R. (2012). Word meaning and concept expressed. *The Linguistic Review*, *29*(4), 607–623.

Chomsky, N. (1980). *Rules and representations.* New York: Columbia University Press.

Chomsky, N. (2000). Minimalist inquiries: The framework. In R. Martin, D. Michaels, & J. Uriagereka (Eds.), *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik* (pp. 89–155). Cambridge: MIT Press.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cover, T., & Thomas, J. (2006). *Elements of information theory*. John Wiley & Sons, Hoboken, NJ.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, *2*, 223–235.

Cronk, B. C., Lima, S. D., & Schweigert, W. A. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, *22*(1), 59–82.

Cronk, B. C., & Schweigert, W. A. (1992). The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics*, *13*(2), 131–146.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* (pp. 49–66). Cambridge, MA: MIT Press.

De Leeuw, J. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavioral Research Methods*, *47*(1), 1–12.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*, *1*, 4171–4186.

Di Sciullo, A. M., & Williams, E. (1987). *On the definition of word*. MIT Press, Cambridge, MA.

Dowty, D. (2007). Compositionality as an empirical problem. In C. Barker & P. I. Jacobson (Eds.), *Direct compositionality* (pp. 14–23). Oxford: Oxford University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Dyer, W., Futrell, R., Liu, Z., & Scontras, G. (2021). Predicting cross-linguistic adjective order with information gain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 957–967).

Espinal, M. T., & Mateu, J. (2010). On classes of idioms and their interpretation. *Journal of Pragmatics*, *42*(5), 1397–1411.

Falkum, I. L., & Vicente, A. (2015). Polysemy: Current perspectives and approaches. *Lingua*, *157*, 1–16.

Fano, R. (1961). *Transmission of information: A statistical theory of communications*. Cambridge, MA: MIT Press.

Fazly, A., Cook, P., & Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, *35*, 61–103.

Fazly, A., & Stevenson, S. (2006). Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL'06)* (pp. 337–344). Trento, Italy.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review*, *100*(2), 233-253.

Ferreira, J., & Pereira Lopes, G. (1999). A local maxima method and a fair dispersion normalization for extracting multiword units from corpora. In *Proceedings of the 6th Meeting on the Mathematics of Language* (pp. 369–381).

Finn, C., & Lizier, J. T. (2018). Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy*, *20*(4), 297.

Fodor, J., & Katz, J. (1964). *The structure of language: Readings in the philosophy of language*. Englewood Cliffs, N.J.: Prentice-Hall.

Frege, G. (1923). Gedankengefüge. In P. Geach & R. H. Stoothof (Eds.), *Logical Investigations. Gottlob Frege, 1977* (pp. 55–78). Oxford: Blackwell.

Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass*, *3*(1), 111–127.

Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019, August). Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the 5th International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)* (pp. 3–13). Paris, France.

Gazdar, G. (1981). Unbounded dependencies and coordinate structure. *Linguistic Inquiry*, *12*(2), 155–184.

Gazdar, G., Klein, E., Pullum, G., & Sag, I. A. (1985). *Generalized phrase structure grammar*. Cambridge, MA: Harvard University Press.

Gibbs, R. W., Nayak, N. P., & Cutting, J. C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, *28(5)*, 576–593.

Glasbey, S. (2007). Aspectual composition in idioms. In L. de Saussure, J. Moeschler, & G. Puskás (Eds.), *Recent Advances in the Syntax and Semantics of Tense, Aspect and Modality* (pp. 71–88). Berlin, New York: De Gruyter Mouton.

Goldsmith, J. A., & Riggle, J. (2012). Information theoretic approaches to phonological structure: The case of Finnish vowel harmony. *Natural Language & Linguistic Theory*, *30*, 859-896.

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

Gries, S. T. (2019). Chapter 2: Polysemy. In E. Dąbrowska & D. Divjak (Eds.), *Cognitive Linguistics - Key Topics* (pp. 23–43). Berlin, Boston: De Gruyter Mouton.

Groenendijk, J., & Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and Philosophy*, *14*(1), 39–100.

Gutknecht, A. J., Wibral, M., & Makkeh, A. (2021). Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A*, *477*.

Hahn, M., Degen, J., & Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, *128*(4), 726–756.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *9117*(5), 2347–2353.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Meeting of the North American chapter of the Association for Computational Linguistics and Language Technologies* (pp. 1–8). Pittsburgh, PA: Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Hay, J. (2003). *Causes and consequences of word structure.* New York, NY: Routledge.

Hendriks, H. (2001). Compositionality and model-theoretic interpretation. *Journal of Logic, Language, and Information*, *10*(1), 29–48.

Hodges, W. (2001). Formal features of compositionality. *Journal of Logic, Language, and Information*, *10*(1), 7–28.

Hoover, J. (2024). *The cost of information: Looking beyond predictability in language processing.* (Doctoral thesis, McGill University)

Husserl, E. (1900). *Logische untersuchungen II/1, translated by J. N. Findlay as Logical Investigations, 1970.* London and Henley: Routledge and Kegan Paul.

Jackendoff, R. (2002). What's in the lexicon? In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.), *Storage and Computation in the Language Faculty.* Dordrecht: Kluwer Academic Press.

Janssen, T. (1986). *Foundations and applications of Montague grammar.* Amsterdam: Centre for Mathematics and Computer Science.

Janssen, T., & Partee, B. (1997). Chapter 7 - compositionality. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of logic and language* (p. 417-473). Amsterdam: North-Holland.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Katz, G., & Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (pp. 12–19). Sydney, Australia: Association for Computational Linguistics.

Katz, J. (1972). *Semantic theory.* New York: Harper & Row.

Katz, J., & Postal, P. (1963). Semantic interpretation of idioms and sentences containing them. In *Quarterly Progress Report of the MIT Research Laboratory of Electronics 70* (pp. 275–282).

Kayne, R. (1994). *The antisymmetry of syntax.* Cambridge, MA: MIT Press.

Kazmi, A., & Pelletier, F. (1998). Is compositionality formally vacuous? *Linguistics and Philosophy*, *21(6)*, 629–633.

Keine, S. (2013). *On idioms and compositionality.* (Handout)

Kintsch, W. (2000). A computational theory of metaphor comprehension. *Psychonomic Bulletin & Review*, *7*, 257–266.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1534–1543.

Kolchinsky, A. (2022). A novel approach to the partial information decomposition. *Entropy*, *24*(3), 403.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Kyriacou, M., Conklin, K., & Thompson, D. (2021). When the idiom advantage comes up short: Eye tracking canonical and modified idioms. *Frontiers in Psychology*, *12*.

Lahav, R. (1989). Against compositionality: The case of adjectives. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *57*(3), 261–279.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Volume 1: Theoretical prerequisites.* Stanford: Stanford University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (pp. 2231–2234). Genoa: European Language Resources Association (ELRA).

Libben, M. R., & Titone, D. (2008). The multidimensional nature of idiom processing. *Memory & Cognition*, *36(6)*, 1103–1121.

List, J.-M., & Forkel, R. (2021). *A Python library for historical linguistics. Version 2.6.9.*

Lovseth, K., de la Parra, L., Wagner, M., & Titone, D. (2011). Familiarity and decomposability modulate the prosodic realization of figuratively vs. literally intended idioms during natural speech production. In D. Ostry, S. R. Baum, L. Ménard, & V. L. Gracco (Eds.), *Proceedings of the 9th International Seminar on Speech Production* (pp. 377–384).

Makkeh, A., Gutknecht, A. J., & Wibral, M. (2021). Introducing a differentiable measure of pointwise shared information. *Physical Review E*, *103*(3), 032149.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (pp. 55–60).

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech 2017* (pp. 498–502).

McGinnis, M. (2002). On the systematic aspect of idioms. *Linguistic Inquiry*, *33*(4), 665–672.

Montague, R. (1970). Universal grammar. *Theoria*, *36*(3), 373–398.

Montague, R. (1974). Universal grammar. In *Formal philosophy: Selected papers by Richard Montague* (pp. 222–246). New Haven, CT: Yale University Press.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*(3), 491–538.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. Cambridge, MA: MIT Press.

O'Grady, W. (1998). The syntax of idioms. *Natural Language & Linguistic Theory*, *16*(2), 279–312.

Pagin, P., & Westerståhl, D. (2010a). Compositionality I: Definitions and variants. *Philosophy Compass*, *5*(3), 250–264.

Pagin, P., & Westerståhl, D. (2010b). Compositionality II: Arguments and problems. *Philosophy Compass*, *5*(3), 265–282.

Palmer, F. R. (1981). *Semantics*. Cambridge, New York: Cambridge University Press.

Partee, B. (1984). Compositionality. *Varieties of Formal Semantics*, *3*, 281–311.

Pelletier, F. (2000). Semantic compositionality: Free algebras and the argument from ambiguity. In M. Faller, S. Kaufmann, & M. Pauly (Eds.), *Formalizing the Dynamics of Information* (pp. 207–218). Stanford: CSLI Publications.

Pelletier, F. (2004). The principle of semantic compositionality. In S. Davis & B. Gillon (Eds.), *Semantics: A Reader* (pp. 133–156). New York: Oxford University Press.

Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., & Cotterell, R. (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* (pp. 949–962). Association for Computational Linguistics.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193.

Portner, P. (2005). *What is meaning? Fundamentals of formal semantics*. Malden, Oxford, Victoria: Blackwell.

Pulman, S. G. (1993). The recognition and interpretation of idioms. In C. Cacciari & P. Tabossi (Eds.), *Idioms—Processing, Structure, and Interpretation* (pp. 249–270). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Radford, A. (2004). *English syntax: An introduction*. Cambridge: Cambridge University Press.

Rambelli, G., Chersoni, E., Senaldi, M. S. G., Blache, P., & Lenci, A. (2023). Are frequent phrases directly retrieved like idioms? An investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)* (pp. 87–98). Association for Computational Linguistics.

Rathi, N., Hahn, M., & Futrell, R. (2021). An information-theoretic characterization of morphological fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10115–10120). Association for Computational Linguistics.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*(1-2), 127–159.

Riehemann, S. Z. (2001). *A constructional approach to idioms and word formation*. (Doctoral thesis, Stanford University)

Rothstein, S. (1991). Heads, projections, and categorial determination. In K. Leffel & D. Bouchard (Eds.), *Views on Phrase Structure* (pp. 97–112). Dordrecht: Kluwer.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science* (pp. 1–15). Berlin, Heidelberg: Springer.

Sag, I. A., & Wasow, T. (1999). *Syntactic theory: A formal introduction.* Stanford: CLSI Publications.

Sandra, D. (1998). What linguists can and can't tell you about the human mind: A reply to Croft. *Cognitive linguistics*, *9*(4), 361–378.

Sa-Pereira, F. (2016). *Distributional representations of idioms.* (Master's thesis, McGill University)

Schweigert, W. A. (1986). The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research*, *15*, 33–45.

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

Shutova, E., Van de Cruys, T., & Korhonen, A. (2012). Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of COLING 2012: Posters* (pp. 1121–1130).

Siyanova-Chanturia, A., & Lin, P. M. S. (2018). Production of ambiguous idioms in English: A reading aloud study. *Journal of Applied Linguistics*, *28*(1), 58–70.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Socolof, M., Cheung, J., Wagner, M., & O'Donnell, T. (2022). Characterizing idioms: Conventionality and contingency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4024–4037). Dublin, Ireland: Association for Computational Linguistics.

Sporleder, C., & Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL* (pp. 754–762). Athens, Greece: Association for Computational Linguistics.

Sporleder, C., & Li, L. (2014). Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (p. 2019-2027).

Stone, M. S. (2016). *The difference between bucket-kicking and kicking the bucket: Understanding idiom flexibility.* (Doctoral thesis, University of Arizona)

Studený, M., & Vejnarová, J. (1998). The multi-information function as a tool for measuring stochastic dependence. In *Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models.* Norwell, MA: Kluwer Academic Publishers.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* (Vol. 54). Cambridge University Press.

Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 523–534.

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

Szabó, Z. (2000). Compositionality as supervenience. *Linguistics & Philosophy*, *23*, 475–505.

Szabó, Z. G. (2004). Compositionality. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.). Stanford University: Metaphysics Research Lab.

Tabossi, P., Fanari, R., & Wolf, K. (2008). Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 313–327.

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition*, *37*(4), 529–540.

Tabossi, P., Wolf, K., & Koterle, S. (2009). Idiom syntax: Idiosyncratic or principled? *Journal of Memory and Language*, *61*(1), 77–96.

Taylor, J. R. (2003). *Linguistic categorization.* Oxford: Oxford University Press.

Taylor, J. R. (2008). Polysemy and the lexicon. In G. Kristiansen, M. Achard, R. Dirven, & F. J. R. de Mendoza Ibáñez (Eds.), *Cognitive Linguistics* (pp. 51–80). De Gruyter Mouton.

Titone, D., & Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, *31*(12), 1655–1674.

Titone, D., & Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon*, *9*(3), 473-496.

Titone, D., Lovseth, K., Kasparian, K., & Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? Evidence from eye movement measures of reading. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *73*(4), 216–230.

Tremblay, A., & Baayen, H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 151–173). London: The Continuum International Publishing Group.

Tuggy, D. (1999). Linguistic evidence for polysemy in the mind: A response to William Croft and Dominiek Sandra. *Cognitive linguistics*, *10*(4), 343–368.

Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, *35*(2), 251–296.

van de Cruys, T. (2011). Two multivariate generalizations of pointwise mutual information. In C. Biemann & E. Giesbrecht (Eds.), *Proceedings of the Workshop on Distributional Semantics and Compositionality* (pp. 16–20). Stroudsburg, PA: Association for Computational Linguistics.

van der Linden, H. J. B. M. (1992). Incremental processing and the hierarchical lexicon. *Computational Linguistics*, *36*, 219–238.

van Lancker, D., Canter, G. J., & Terbeek, D. (1981). Disambiguation of ditropic sentences: Acoustic and phonetic cues. *Journal of Speech & Hearing Research*, *24*(3), 330–335.

Vicente, A., & Falkum, I. L. (2017). Polysemy. In M. Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics.* New York: Oxford University Press.

von Humboldt, W. (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* (pp. 401–430).

Wagner, M. (2021). *Prosodylab experimenter.*

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IMB Journal of Research and Development*, *4*(1), 66–82.

Weinreich, U. (1969). Problems in the analysis of idioms. In J. Puhvel (Ed.), *Substance and Structure of Language* (pp. 23–81). University of California, Los Angeles: University of California Press.

Westerståhl, D. (2002). On the compositionality of idioms. In *Proceedings of LLC8.* CSLI Publications.

Westerståhl, D. (1998). On mathematical proofs of the vacuity of compositionality. *Linguistics and Philosophy*, *21*(6), 635–643.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023, 12). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470.

Williams, A., Cotterell, R., Wolf-Sonkin, L., Blasi, D., & Wallach, H. (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.

Williams, P., & Beer, R. (2010). *Nonnegative decomposition of multivariate information.* arXiv preprint arXiv:1004.2515.

Wollstadt, P., Lizier, J. T., Vicente, R., Finn, C., Martinez-Zarzuela, M., Mediano, P., ... Wibral, M. (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.

Wood, M. M. (1986). *A definition of idiom*. Indiana University Linguistics Club.

Wu, S., Cotterell, R., & O'Donnell, T. J. (2019). Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5117–5126). Florence, Italy: Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, 5754–5764.

Zadrozny, W. (1994). Free compositional to systematic semantics. *Linguistics and Philosophy*, *14*(4), 329–342.