DECOMPOSING IDIOMS: FACTORS THAT IMPACT PERCEIVED IDIOMATICITY

by

KATHRYN CONGER

B.A., University of Colorado, 2010

M.A., University of Colorado, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Linguistics and Institute of Cognitive Science

2025

Committee Members:

Bhuvana Narasimhan

Al Kim

Laura Michaelis

Martha Palmer

James Martin

ABSTRACT

Conger, Kathryn Summerville (Ph.D., Linguistics and Cognitive Science)
Decomposing Idioms: Factors that Impact Perceived Idiomaticity
Thesis directed by Professor Bhuvana Narasimhan & Associate Professor Al Kim

Though prevalent in everyday language, idioms are notoriously difficult to define. This dissertation investigates factors that impact our conceptualization of idioms by addressing the challenge of defining idiomaticity as a cognitive construct. To this end, it contributes the first empirically tested description of the class idiom. Idioms are often operationalized as though they comprise a cleanly delineated class, in which a phrase either is, or is not, idiomatic based on the necessary and sufficient condition of noncompositionality. This is carried into psycholinguistic work investigating idiom processing and serves as the basis for several prevalent models of idiom comprehension. However, some propose that idioms exist on a continuum, with certain properties of idiomaticity moderating the degree to which phrases are perceived as idiomatic. Experiment 1 presents direct evidence demonstrating that idiomaticity is not a discrete construct, thereby resolving this point of contention. Instead, it shows that the mental category idiom is continuous, and membership may be based on prototype similarity. Experiment 2 investigates the relationship between perceived idiomaticity and several properties of idioms, finding that distributed idiomatic meaning, partial literality, and plausible dual idiomatic and non-idiomatic meaning are individually and collectively predictive of perceived idiomaticity and may serve as markers of idiom prototypicality. Overall, this work increases our understanding of the mental construct of idiomaticity, addressing long-standing assumptions about the mental representation of phrases and the access of idiomatic meaning, upon which models of idiom conceptualization and

comprehension are predicated. Additionally, it raises methodological and theoretical concerns, urging for careful consideration of properties in experimental design as well as in the interpretation of experimental findings. More broadly, these findings have implications for theories of language processing, second language acquisition, natural language processing, and clinical linguistics. Thus, this study paves the way for a more nuanced, prototype-based account of idiomaticity, capable of addressing variation within the class idiom by recognizing and incorporating the continuum along which idiomatic phrases lie.

DEDICATION

To my mom, Sue Conger, whose support, encouragement, and patience while suffering through countless tedious discussions on idioms made this possible.

ACKNOWLEDGEMENTS

I extend my sincere gratitude to everyone who made this dissertation and this degree possible, both academically and personally.

In particular, a huge thank you to my advisors, Bhuvana Narasimhan and Al Kim, for your guidance, feedback, and patience with endless conversations about idioms. I am so grateful to Bhuvana for her trust, support, and willingness to allow me to pursue this topic, no matter what. I am very grateful to Al for the many insightful discussions and lab meetings that broadened my perspective on language and introduced me to pointwise mutual information.

I am also grateful to my committee members for their time, engagement, and guidance. Laura Michaelis introduced me to the wonderful world of idioms during my first semester in the PhD program. Her love of the topic was infectious, and it wasn't long before my planned topic of study was replaced by idioms. They never cease to amaze, and I couldn't be more grateful. I have never looked back. Martha Palmer has been part of my academic path since long before this dissertation. Her guidance and mentorship on all fronts have been invaluable. So too have the countless opportunities I have been given, experiences I have had and the linguistic training I received. While Laura ignited my love of idioms, Martha made a PhD possible. Jim Martin's comments and guidance connected interdisciplinary dots. His feedback shifted crucial puzzle pieces so that they fit together smoothly and created a complete picture (at least in my own head).

I extend my sincere appreciation and gratitude to friends and to members of GRoW for their support and encouragement along the way, not to mention the multiple rounds of task development and piloting you gracefully suffered through. I am especially thankful to Jennifer Fitzgerald, who offered crucial support and guidance

at key moments during this process. Her generosity and encouragement made it possible to continue forward.

To my mom, thank you for your endless encouragement, engagement, and patience at every stage of this project. You're closing in on an honorary degree in idioms, whether you wanted one or not.

Finally, I gratefully acknowledge the support of the Linguistics Department, ICS, and CARTSS for funding this work. Their support made it possible to conduct pilot studies, run experiments, and bring this work to completion. I am also extremely grateful to instructors who allowed me not only to recruit participants from their classes but also offered extra credit in exchange for participation. In particular, Dr. Rai Farrelly, whose help over multiple semesters allowed for a sample that was balanced between university students and Prolific participants across all tasks.

CONTENTS

TABLES

Table

FIGURES

Figure

CHAPTER I

INTRODUCTION

When we read or hear phrases such as *break the vase*, we analyze words one at a time, integrating the meaning of each into an unfolding sentential representation (cf. Hagoort 2005). But, what happens if the meaning of a phrase is not the sum of its parts? How is it that your brain willingly accepts *break the ice* to mean something like "facilitate social interaction by alleviating awkwardness" rather than to actually "break frozen water" when most native English speakers would not accept a phrase such as *break the vase* as meaning anything other than to actually "break a vase"? Is it because *break the ice* is known as an idiom while *break the vase* is not? Or, is it because of something more general, such as knowledge of how words in the phrase work together to create meaning?

The goal of this work is to investigate factors that impact our conceptualization of idioms by addressing the challenge of defining the cognitive construct of idiomaticity to unpack what makes a phrase recognizable as an idiom. It is estimated that there are about 25,000 idioms in American English, roughly as many as the number of individual words the average adult speaker knows (Thyab 2016). Given their prevalence, we can conclude that native speakers are familiar with idiomatic phrases. However, whether they recognize a particular subset of words as idiomatic is not clear. Moreover, the definition of an idiom is elusive, leaving a gap in our understanding of how we think about and understand idioms. These omissions partially motivate this work.

The most commonly cited definition of an idiom is a phrase whose meaning cannot be derived by combining the individual meanings of its component parts (Hockett 1958, Katz 1973, Weinreich 1969, Pulman 1993, Marlies 1995, Mel'čuk 1995)[1]. While the meaning of most non-idiomatic phrases is *ordinarily denoting*, in that the meaning denotes the same category of items within the phrase as it does outside of it, all idioms have *non-compositionally denoted* meaning, because the meaning of the phrase *is not* the denotational sum of the component parts (Ifill 2019, see also Nunberg et al. 1994, Espinal & Mateu 2019). For example, *cold turkey* does not refer to an actual turkey with the feature [+cold] but instead to "the abrupt cessation of a habitual activity" (Pitt & Katz 2000).

While it is true that non-compositionality is a hallmark of idiomaticity, this single criterion is not sufficient to differentiate idioms from other types of phrases (e.g., sarcastic statements). Additionally, it homogenizes phrases within the category idiom, assuming that noncompositionality is unique to idioms and that all noncompositional phrases are equally noncompositional. This leads to a discussion of idioms as if they form a discrete class, cleanly separated from non-idioms. However, noncompositional meaning is not unique to these phrases.

Additionally, noncompositionality is not the only characteristic shared by these phrases. For example, because the meaning of an idiom is not the sum of its parts,

[1]Note that this semantically based definition differs from the syntactic definition used in Sign-Based Construction Grammar (SBCG), which states that an idiom is a phrase whose combinatorics are not permitted by canonical phrase-structure rules (Michaelis 2013). Further, in a constructional idiom, at least one slot is lexically fixed (Booij 2013). However, phrases vary in their degree of lexical fixity, which forms a continuum (Croft & Cruse 2004, Kay & Michaelis 2012; Michaelis 2012, 2013, 2017, 2019). At the least flexible end are completely frozen idioms that have idiosyncratic syntax (e.g., by and large). At the opposite end of the continuum are productive constructions where the pattern itself is set but any word can fill the slot, such as in canonical phrases, imperatives, and questions. While this work focuses on semantics, it operates at the syntax-semantics interface. The idioms and non-idiomatic collocations included in this work are situated between the fixed and semi-flexible points on the continuum of lexical fixity. The work should be viewed not only as comparable with SBCG but as zooming in on the continuum of lexical fixity.

all idioms are also formulaic. At its base, the term ***formulaic phrase*** refers to a probabilistic aspect of language and can be defined as a pattern of words that appear together more often than would be expected by chance. These highly frequent patterns are recognizable and predictable to native speakers. The implications of formulaicity extend to semantics and syntax - speakers have to know the order in which words must appear to evoke an idiomatic interpretation of a phrase because the sequence in which words appear is, itself, meaningful (cf. Wray & Perkins 2000, Wray 2002, Sanchez-Lopez 2015, Carrol & Conklin 2019).

In addition to formulaicity, a number of linguistic and psycholinguistic properties are associated with idiomaticity. These include partial literality, dual meaning, distributed idiomatic meaning, informality, affect, proverbiality, figuration, inflexibility, and transparency (Nunberg et al. 1994:492-493, Fernando 1996:3, Barkema 1996:128, Penttilla 2010). Of these, partial literality, dual meaning, and distributed meaning may be particularly important for differentiation and characterization of idiomaticity as a cognitive class, that is, for determining what makes a phrase seem idiomatic. For this reason, they will be introduced last.

Informality and affect are contextual dimensions with informality referring to the register in which idioms are most commonly found and affect referring to the evaluative stance conveyed, since idioms are frequently used in emotionally charged social situations (Nunberg et al. 1994). Proverbiality refers to a subset of idioms that are common cultural sayings. Such sayings describe recurring social situations and often impart warnings or advice such as *the early bird gets the worm* or *a bird in hand is worth two in the bush*. While linguistically important, little attention is paid to these dimensions in the psycholinguistic literature.

Figuration refers to whether an idiom is based on a figurative mapping. These phrases may be metaphoric, metonymic, or hyperbolic, among other tropes (Nunberg et al. 1994). Inflexibility refers to the generally fixed nature of idioms. The degree to which a phrase is inflexible is assessed by determining whether the component parts of an idiom may be modified or changed without losing idiomatic phrasal meaning (Pulman 1986). For example, some phrases allow for internal modification (1a) or constituent movement (2a) while others do not (1a and 2b).

(1a.) *Leave no **legal** stone unturned.*[2]

(1b.) **by and **very** large*

(2a.) ***Those strings**, he wouldn't pull for you.*[3]

(2b.) ***The nose**, people will pay through.*

Transparency refers to whether there is a clear rationale for the figurative meaning of an idiom (Nunberg et al. 1994). For example, *saw logs* and *pop the question* are transparent because there is a clear relationship between their literal[4] and idiomatic meaning. *Saw logs* equates the sound of snoring to that of sawing logs while *pop* is metaphorically extended in *pop the question*. On the other hand, *spill the beans* and *kick the bucket* are opaque because there is no obvious relationship between their literal and idiomatic interpretations. While more attention has been paid to inflexibility, transparency, and figuration than informality, affect, and

[2] portal.ct.gov

[3] Chae 2015, p.50s

[4] At times, "literal" is used as the opposite of "idiomatic". This was not taken lightly but was unavoidable due to the way in which certain constructs have historically been discussed and operationalized. The line between "literal" and "non-literal" is just as fuzzy as the line between "idiom" and "non-idiom". Additionally, the category "non-idiom" includes all types of language except idioms. The category "literal" is one type "non-idiom", however there are many other category members (e.g., metaphors and other tropes). However, in the psycholinguistic literature, the use of "literal" as the antonym of "idiomatic" is ubiquitous. To stay true to the literature and to avoid possible misinterpretations when the use of "literal" is intended to refer to truly literal phrases (e.g., *red ball*) versus non-idiomatic phrases that may not be literal (e.g., *red head*), "literal" will be used. A full discussion of inconsistent and problematic usage of "literal" in the idiom comprehension literature, and elsewhere, is outside of the scope of this work.

proverbiality, there is little evidence that these properties independently impact idiom comprehension. For example, Gibbs & Nayak (1989) found that idiom flexibility impacts native speakers' judgements of idioms in online and offline tasks. However, this effect was only significant when distributed idiomatic meaning, which refers to whether idiomatic meaning is associated with individual words (see below for discussion) was manipulated, demonstrating that flexibility alone does not seem to impact idiom processing[5].

Partial literality, dual meaning, and distributed idiomatic meaning play a central role in this work. ***Partial literality*** refers to whether at least one word in an idiomatic phrase contributes the same meaning it has outside of the phrase (Gibbs & Nayak 1989; Titone & Connine 1994a, 1999; Libben & Titone 2008, Nordmann et al. 2014). For example, *foot the bill,* is partially literal because *bill* refers to "an invoice" both within this phrase and outside of it. However, *kick the bucket* is not partially literal because neither *kick* nor *the bucket* contributes the meaning that these words carry outside of the phrase.

Dual meaning refers to whether a phrase can be used both idiomatically and non-idiomatically or only idiomatically (Gibbs & Nayak 1989, Titone & Connine 1994a, 199b, 1999, 2014, Libben & Titone 2008, 2011, Cailles & Butcher 2007; Nordmann et al. 2014). For example, *bad apple* has dual meaning because it can be used idiomatically to refer to "a bad or corrupt person in a group, typically one whose behavior is likely to have a detrimental influence on their associates" (oed.com) and

[5] This is not to say that flexibility is not important. Instead, it highlights the multidimensional nature of idiomaticity, arguing for careful consideration not only of the individual impact of idiom properties but also relationships between properties and their combined impact on idiom conceptualization and comprehension. See section 2.2.2 for further discussion on the relationship between distributed idiomatic meaning and flexibility.

non-idiomatically to refer to "a rotten apple". However, *funny bone* can be used idiomatically only since there is no actual "bone that is funny".

Distributed idiomatic meaning refers to whether individual words in an idiomatic phrase contribute meaning or not (Gibbs & Nayak 1989, Gibbs et al. 1989, Nunberg et al. 1994, Titone & Connine 1994, Tabossi et al. 2008, Nordmann et al. 2014, Libben & Titone 2014). For example, *spill the beans* ("to unintentionally, prematurely, or indiscreetly reveal secret or privileged information" (oed.com)) has distributed idiomatic meaning because *spill* contributes the meaning of "to reveal" and *the beans* contributes the meaning of "information". However, *kick the bucket* ("to die") does not have distributed idiomatic meaning because the meaning of "to die" is not contributed by either *kick* or *the bucket*.

The definition of distributed idiomatic meaning is more nuanced than the others and therefore deserves further clarification. Historically, this has been referred to as *decompositionality*. Taking a step back, compositionality refers to a parent class which includes three subtypes of denoted meaning: normal, nondecompositional and *decompositional* (Nunberg et al. 1994, Pitt & Katz 2000, Cacciari 2014). The denoted meaning of a phrase is said to be normally compositional when its meaning is the sum of the parts (e.g., *purple car* = "a vehicle that is a color created by combining red and blue"). By definition, idioms are noncompositional, meaning that they are not normally compositional as their meaning is not the sum of the literal meaning of the parts. Within psycholinguistics, those who adopt a categoric stance use idiomaticity synonymously with noncompositionality (Bobrow & Bell 1973, Swinney & Cutler 1979, Qualls & Harris 2018). However, noncompositional idioms can be further differentiated based on distributed idiomatic meaning. Nondecompositional idioms are idioms with no distributed idiomatic meaning, such as *kick the bucket*. In this phrase, neither *kick* nor *bucket* contributes the phrasal

meaning of "to die". Decompositional idioms, such as *spill the beans*, have distributed idiomatic meaning. *Spill the beans* is not normally compositional because its meaning is not created by combining the meaning that the individual parts carry outside of the idiomatic phrase (e.g., *spill the beans* ≠ "knock over the legumes"). *Spill the beans* differs from nondecompositional idioms in that idiomatic meaning is contributed by each constituent (e.g., *spill* = "reveal" *the beans* = "information". As an alternative to those who take a categoric approach, others take a continuous approach, positing that differences in distributed idiomatic meaning impact perceived idiomaticity and comprehension (Gibbs & Nayak 1989; Gibbs et al. 1989; Titone & Connine 1994b, 1999; Gibbs et al. 1997, Caillies & Butcher 2007; Libben & Titone 2008, Titone & Libben 2014, Titone et al. 2019).

Given the number of properties associated with idioms, it is not surprising that there is a great degree of variation within the class, with some phrases seemingly more similar to non-idioms than other idioms. Thus, treatment of idioms as a discrete class may be problematic as there is little reason to assume that there is a binary distinction between a cleanly defined class "idiom" and a separate class "non-idiom". Within linguistics, many have advocated for a prototype-based account of idioms, with more prototypically idiomatic phrases and prototypically non-idiomatic phrases at opposite ends of a continuum and less prototypical members of each class situated closer to the middle (cf. Rosch 1973, Rosch & Mervis 1975, Kamp & Partee 1995, Geeraerts 1989, Wulff 2008, Pentilla 2010, Watson 2019, Rosch & Lloyd 2024. For example, prototypically collocative, non-idiomatic phrases such as *government spending* and *heavy rain* would exist at one end of the continuum, representing phrases that are the furthest from the prototypical idiom. Although the meaning of these highly formulaic collocations has been conventionalized, their meaning is easily understood by combining the meanings of the individual constituent words.

However, differences in prototypicality and, by extension, their position on the continuum, are evident as both words in *government spending* are used in a literal sense while *heavy* in *heavy rain* is not entirely literal as it does not refer to "rain that weighs a lot" but metaphorically conveys information about the density of the rain[6]. At the other end of the spectrum are prototypically idiomatic phrases, such as *sweet tooth* or *grease monkey*. The meaning of a prototypical idiom cannot be decomposed and assigned to the individual constituents. Additionally, it is generally unrelated to the combined meaning of the individual constituent words although, some clues may remain. For example, neither *sweet* nor *tooth* contribute identifiable idiomatic or non-idiomatic meaning to the phrase, making the phrase relatively opaque. However, *sweet* establishes a reference domain of "sweet items" leading to an association between *tooth* "eating". While this phrase would be opaque to a non-native speaker, native speakers share tacit experiential knowledge about this phrase based on prior experience (Wulff 2024). Finally, peripheral collocations, such as *strong coffee* and peripheral idioms, such as *blanket statement* exist toward the

[6] In American English, *heavy* is commonly used metaphorically to describe intensity or volume, not just physical weight. By contrast, *dense* is used to describe things that are tightly packed or thick (e.g., *dense fog, dense crowd*). For example, *heavy rain* implies a large volume of rain falling intensely, which aligns with how we experience rainstorms. The rain itself is not perceived as a "thick" substance in the same way fog might be. In the case of *dense crowd, dense* is associated with "tightly packed". When "tightly packed" entities are in a contained area, creating substantial crowding or even filling the space entirely, they are perceived as a "thick substance". This occurs because substantial crowding results in an inability to individuate entities within the crowd. Interestingly, *dense fog* is perceived of as less idiomatic than *heavy rain,* even though neither is an idiom and both adjectives are used metaphorically. This may be because the first dictionary entry for adjectival *dense* is for the sense conveyed by *dense fog* (https://www.thefreedictionary.com/dense) while the most common meaning of *heavy* is "having relatively great weight' (https://www.thefreedictionary.com/heavy). A similar example can be seen with *blanket statement,* which most native speakers view as an idiom (Conger 2022). While it is *blanket* that triggers this perception, this word conveys a meaning inside the phrase that it can outside of the phrase. The two senses of adjectival blanket are: "1. Applying to or covering all conditions or instances: *a blanket insurance policy*, and 2. Applying to or covering all members of a class: *blanket sanctions against human-rights violators*." (https://www.thefreedictionary.com/blanket). It is for reasons such as these that we might expect *heavy rain* and *government spending* to fall on different points on a continuum. Such findings also indicate the need for a careful consideration of figuration in the study of non-idiomatic collocations (for the role of figuration in idiom processing, see section 3.2.2).

middle of the continuum. Peripheral phrases tend to be less syntactically flexible than strong collocations but more flexible than strong idioms.

Within psycholinguistics, clear delineation of the class idiom is a longstanding point of contention. With respect to how we think about and understand idioms, some researchers take a discrete category approach, theorizing that all members of the class idiom are activated via an idiom-specific processing strategy when a phrase is recognized as an idiom (cf. Bobrow & Bell 1973, Swinney & Cutler 1979, see also Cacciari & Tabossi 1988, Canal & Bambini 2023). From this viewpoint, mental stipulation of a phrase as idiomatic is a requisite for efficient processing. Other researchers posit that similarities between idioms and other types of language, as well as differences within the class idiom, render a "one-size-fits-all" treatment impossible (cf. Gibbs & Nayak 1989; Gibbs et al. 1989; Titone & Connine 1994b, 1999; Gibbs et al. 1997; Caillies & Butcher 2007; Libben & Titone 2008; Titone & Libben 2014). Instead, idioms exist on a continuum. From this viewpoint, properties associated with idioms impact *perceived idiomaticity*, or one's perception of a phrase as idiomatic or non-idiomatic, moderating recognition and conceptualization of idiomatic phrases. Both groups use the same, but differently interpreted, experimental findings to support their conclusions, and these same findings will continue to be used to substantiate both sides of this debate until critical knowledge gaps are addressed.

The goal of this research is to address two foundational points underlying theories of how idiomaticity and idiom properties impact idiom conceptualization: the assumption that idioms form a discrete class such that idiomaticity, not idiom properties, account for the perception of a phrase as a member of the class idiom, and the assumption that multiple properties impact conceptualization, moderating

the degree to which a phrase is idiomatic[7]. To address the first assumption, this research investigates the role of idiomaticity by assessing the validity of a discrete mental category idiom. This is done by determining whether native English speakers are able to reliably differentiate idioms from equally formulaic non-idiomatic collocations, a point central to the claim that all members of the class idiom are mentally handled in the same manner. Next, this research considers the impact of differences between phrases within the class idiom, by investigating the relationship between the idiom properties of formulaicity, distributed idiomatic meaning, partial literality, and dual meaning on perceived idiomaticity in an effort to create a more complete and nuanced understanding of the mental category "idiom". In addition to addressing two long-standing assumptions, this work challenges traditional definitions of an idiom in which the construct can constrain a meaning that is not the sum of the parts (Shenk 1995). Instead, it highlights the multidimensional nature of idiomaticity, laying the groundwork for a prototype-based approach in which idiomaticity is associated with a number of semantic, syntactic, pragmatic, and psycholinguistic properties which can be quantifiably shown to account for perceived idiomaticity.

Ultimately, this research provides a more nuanced understanding of how idioms are characterized, differentiated, and conceptualized, laying the groundwork for a prototype-based approach to idiomaticity. Such insights will contribute to the field of psycholinguistics by addressing long-standing assumptions about idiomatic

[7] This may not seem inherently problematic. However, while individual analysis proponents share the belief that idiomaticity is not sufficient to account for the class idiom, they differ as to the theorized property or properties thought to impact idiom processing. Such inconsistencies are problematic for replication and conclusion generalizability. Additionally, the shared assumption that "properties" impact idiom processing has licensed prioritization of specific properties favored by individual research groups to the exclusion of others, leading to work that might ignore a property found impactful by a separate group in favor of more carefully controlling for subtypes of the property of interest. Thus, it is vital that a more complete understanding of the role of multiple properties is obtained, as this can be used to inform experimental design decisions so as to avoid nuisance variables while also making work more comparable, increasing generalizability between studies.

language and offering potential explanations for variability in idiom comprehension across individuals. More broadly, this work has implications for natural language processing, theories of L2 acquisition, foreign language instruction, and clinical applications as idiom properties may serve as linguistic biomarkers of a number of disorders impacting linguistic structures at the syntax-semantics interface. For example, idiom properties may be useful for the automatic identification of idioms in natural language. With respect to second language instruction, there is a long-standing debate over whether idioms should be explicitly taught or whether a more implicit learning methodology should be employed, in which learners are exposed to idioms but do not receive explicit instruction (cf. Siyanova-Chanturia & Van Lancker Sidtis 2018). Differences in idiom prototypicality might suggest that learners could benefit from explicit instruction for more prototypical idioms. Finally, differences in idiom conceptualization and comprehension have practical applications in the clinical domain. ***Comprehension*** refers to online processes of activating and retrieving the meaning of an idiom, or how one processes language in real time. ***Conceptualization*** involves forming, organizing, and reflecting on mental representations, including abstract ideas, schemas, and meanings (Casasanto & Lupyan 2015). This includes metalinguistic processes such as conscious analysis as well as categorization of a phrase as a member of the class idiom within the broader context of formulaic language (Langacker 1987). An understanding of idiom conceptualization may be particularly useful in the clinical domain. For example, research on Alzheimer's disease (AD) has shown that certain idioms are preserved much further into disease progression than would be expected (Lindholm & Wray 2011, Bridges & van Lancker Sidtis 2013). Anecdotal evidence suggests a relationship between phrase preservation and the degree of prototypicality, suggesting that idioms may be useful in the diagnosis and treatment of AD. However, a more complete understanding of the relationship

between multiple idiom properties and idiom prototypicality are needed in order to extend this work in a useful manner.

The remainder of this thesis is organized as follows. Chapter 2 establishes a background, briefly introducing the two main approaches to idiom processing, which motivate the need for a clear understanding of idiom conceptualization. For each model, a long-standing underlying assumption is identified and the knowledge gaps associated with each are discussed. It concludes with a presentation of the research questions and predictions, which seek to address these assumptions. Chapter 3 presents experiment 1, a ratings task, which addresses the first research question. Chapter 4 presents the findings of the second experiment, three ratings tasks, which, along with the findings of experiment 1, are used to address the second research question. It concludes with chapter 5, which presents the general discussion, implications, limitations, future directions, and conclusions.

CHAPTER II

BACKGROUND

Within psycholinguistics, the way idiomaticity and certain associated properties are handled often perpetuates assumptions about the categoric or continuous nature of the class idiom. The construct of "idiomaticity" is operationalized in one of two ways: either as a discrete class defined by a single dimension or as a multidimensional class with fuzzy boundaries. Those who adopt a discrete class stance operationalize idiomaticity in terms of "noncompositionality", defining idiomaticity by this single, necessary and sufficient condition and theorizing that it creates a natural boundary between idioms and non-idioms. This viewpoint is associated with, what I term "holistic" models, which share the belief that the noncompositional nature of idioms presents an insurmountable challenge to language-general, word-by-word processing theories. Proponents claim that phrases are holistically represented as long words-with-spaces.

Those who adopt a continuous stance operationalize idiomaticity in a multidimensional manner, positing that idiomaticity is associated with a number of defining properties, each of which may impact perceived idiomaticity and comprehension. This viewpoint is associated with, what I call "individual analysis models", which share the belief that individual words contribute to our understanding of phrasal meaning. Additionally, they posit that properties may be displayed to varying degrees. Modulation of properties impacts how we think about and comprehend idioms and may contribute to their degree of perceived idiomaticity.

These two operationalizations of idiomaticity and the models associated with them serve as the foundation upon which psycholinguistic work on idioms is based. However, the disparate nature of these operationalizations raises questions as to the validity of this foundation. While an understanding of the mental construct of idiomaticity is vital to produce beneficial, substantive, rigorous work able to advance the field, our understanding of idiomaticity, and its relationship to properties, is limited. This limited understanding has led to the development of models based on untested assumptions, which in turn has implications for the advancement of the field.

The remainder of this section introduces the various ways that prior work has discussed idiomaticity and certain properties associated with idioms, illustrating untested, foundational assumptions about the nature of idiomaticity and highlighting how treatment of idiomaticity as categorical versus continuous in nature has shaped psycholinguistic theory. Section 2.1 introduces holistic models of idiom comprehension and the role of noncompositionality which is prioritized to the exclusion of other properties, to the point of being used synonymously with idiomaticity. Next, section 2.2 introduces individual analysis models and four properties of idioms that are prioritized in place of a broad, vaguely defined construct of idiomaticity. It discusses how these properties may individually and collectively modulate perceived idiomaticity, impacting the ability of native speakers to distinguish idioms from other types of phrases.

2.1. Holistic models

Holistic models take a categoric approach to idioms, predicating their theory upon the assumption that idioms are quickly and reliably differentiated from non-idioms because they are noncompositional. There are two major holistic models, the idiom

list hypothesis (Bobrow & Bell 1973) and the lexical representation hypothesis[8] (Swinney & Cutler 1979), Holistic models share the assumption that, because they are noncompositional, idioms form a natural class, and that native speakers share intuitions about members such that they are able to be quickly and reliably differentiated from non-idioms.

Predicated on generative theories which focused on accounting for anomalous meaning, transformational deficiencies, and ill-formedness (Chafe 1968 p.111, Fraser 1970, Culicover 1976, Chomsky 1980), holistic models do not differentiate between decompositional idioms (phrases in which a portion of the meaning of the phrase is associated with each word) and nondecompositional idioms (phrases in which no meaning can be attributed to each word. Within the holistic framework, nondecompositionality and decompositionality are conflated (see Figure 1). Instead, they rely on the more general, higher-level category of phrasal noncompositionality, which refers to a phrase whose meaning is not the sum of its parts, regardless of transparency, encoding, flexibility, etc. Additionally, as having a meaning that is not compositional is the only requisite condition of idiomaticity, idiomaticity is operationalized as noncompositionality with the synonymous use of idiomaticity and noncompositionality leading to claims of a discrete class of phrases differentiated from other phrases by virtue of idiomaticity. Thus, idioms form a discrete class in which all members are equally noncompositional while all non-idioms are equally compositional. The first goal of this research is to test this assumption.

[8] Despite the years since its introduction, the lexical representation hypothesis has remained relevant. Although its popularity dipped after the introduction of individual analysis models, it has recently experienced a resurgence in popularity, particularly amongst those adopting neurophysiological methodologies (cf. Papago 2003, Rommers et al. 2013, Canal & Bambini 2013, Canal et al. 2014, Qualls & Harris 2018).

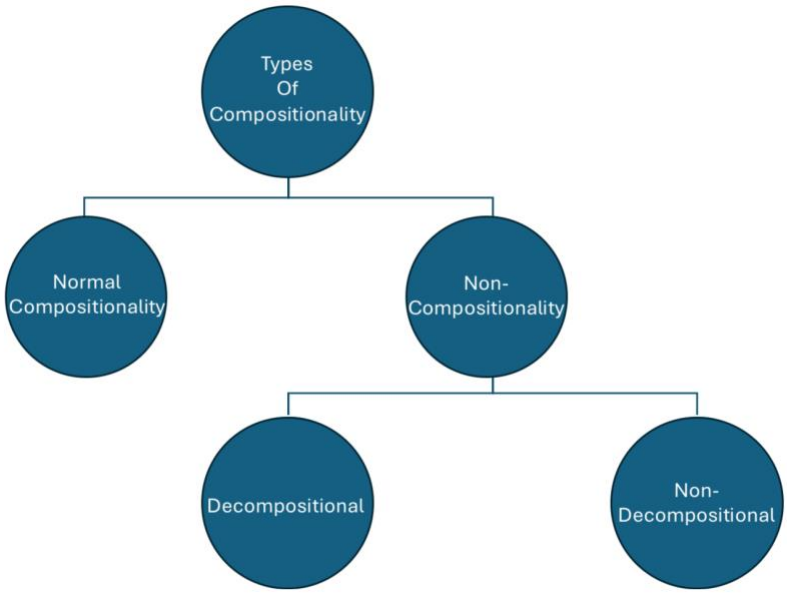


Figure 1. Terminology related to compositionality and noncompositionality.

Holistic models get their name because they believe that all idioms are mentally stipulated as idiomatic and holistically represented as long words-with-spaces in the mental lexicon. Holistic representation is necessary because the literal language processor cannot accommodate noncompositionality. Instead, holistic models posit an idiom-specific strategy in which idiomatic meaning is assessed at the phrasal level only. Meaning is retrieved via a direct mapping between a long idiomatic word and its holistically represented meaning. Crucially, this process is the same for all idioms, homogenizing the class. Thus, idiomaticity, which may be multidimensional, is defined by a single necessary and sufficient condition, that of noncompositionality.

This can be seen with the idiom list hypothesis (Bobrow & Bell 1973). According to Bobrow & Bell (1973), the requisite special treatment of idioms includes an idiom-specific mode of comprehension and representation in which all idioms are

stipulated in the mental lexicon as idiomatic and each is represented as a "word with spaces", or a single conceptual unit, within a speaker's mind. For example, the idiom *break the ice* would be mentally represented as <*break_the_ice*>. During comprehension, individual words in an idiomatic construction are first interpreted literally only. When the literal interpretation does not fit a given sentential context, the phrase is reanalyzed. Lexical-level processing is "switched off" allowing for the phrase to be treated as a single long word. A search is then conducted in a separate lexicon, where long idiomatic words are stored. Here, the meaning of the entire phrase is directly retrieved from the matching holistic mental representation. Thus, recognition of a phrase as an idiom impacts how we think about and understand idioms.

Despite the years since its introduction, holistic models have remained relevant. Although their popularity declined in the 1990s, neurophysiological methodologies sparked a resurgence in their popularity in the 2000s (cf. Papago 2003, Vespignani et al. 2010, Quals et al. 2001, Rommers et al. 2013, Canal & Bambini 2013, Canal et al. 2014, Canal et al. 2017, Canal et al. 2021, Qualls & Harris 2018). One such example comes from Canal et al. (2017), who used event related potentials to investigate whether idiom processing involves an analysis of individual words leading to lexical-level activation or whether idiomatic meaning is holistically represented and retrieved as a single conceptual unit. EEG signals were recorded while participants read sentences containing a literal or idiomatic phrase. They found a pattern of activation for idioms that differed from that observed with literal phrases. While literal phrases showed an expected N400, a P300, but no N400, was observed for idioms. The N400 is a negative deflection of the brain wave that is elicited about 400ms following the presentation of meaningful stimuli and is typically larger when words are less expected (Kutas & Hillyard 1980). Conversely,

the P300 is an earlier component associated with anticipatory mechanisms, such as the pattern matching of words in a pre-stored, long word-with-spaces, which involves a comparison to be sure words in a phrase appear in the expected forms and locations but does not involve processing of their meaning. This finding was interpreted as suggesting that idioms are processed as holistically represented units rather than through the analysis of individual words, supporting the holistic approach to idiom comprehension. Further, the absence of a significant N400 supports the claim that idioms are processed holistically. The N400 is associated with the integration of new information into existing context. A lack of this effect suggests that the meaning of idioms is readily accessed without the need for compositional analysis. These findings imply that idiomatic meaning is retrieved as a single conceptual unit.

2.1.1. Holistic Support and Assumptions

The core tenets of holistic models can be summarized as follows:

> (1) Recognition of a phrase as an idiom as a function of nondecompositionality activates the idiom-specific comprehension strategy.
>
> (2) Failure to recognize a phrase's idiomatic status will result in a processing delay or failure.

These statements rest on a number of assumptions, two of which are central to this work. Behind (1) lies the assumption that idioms form a discrete class by virtue of noncompositionality. Behind (2) lies the assumption that, in most cases, native speakers are able to reliably differentiate idioms from non-idioms. However, there is no direct support for these assumptions because, to date, neither has been directly investigated. Instead, indirect support has come from ***the idiom facilitation effect***, a widely replicated finding demonstrating that idioms are comprehended

more quickly than literal phrases (Swinney & Cutler 1979, McGlone et al. 1994, Tabossi et al. 2009, Rommers et al. 2013).

Holistic models theorize that, if idioms are holistically represented, they should be comprehended more quickly than non-idioms. This is because, once a phrase has been recognized as an idiom, individual words are not analyzed, avoiding homonymy-induced overlapping activation and leading to phrasal meaning retrieval prior to phrase completion. To test this line of reasoning, Swinney & Cutler (1979) presented participants with literal or figurative word strings and measured the time it took them to determine whether word strings were meaningful or not. According to Swinney & Cutler (1979), upon encountering *break*, literal interpretations of <*break*> are activated. Although idiomatic meaning is not actively pursued, *break* is retained for comparison to longer words, as it matches the beginning sounds of other longer words, such as *break_the_ice.* Ambiguity resolution occurs when the phrase is recognized as idiomatic. For example, if the phrase is recognized as idiomatic on the final word, *ice*, the phrase resolves in favor of the idiom. However, should the phrase end with a different word, such as *cup* in *break the* ***cup***, the idiomatic match would no longer be considered for meaning retrieval. When a phrase is resolved prior to the final word in the phrase, idiomatic meaning should be retrieved sooner than literal meaning, simply because remaining words do not undergo individual analysis. Importantly, in order for this transition from literal processing to holistic idiom processing to occur, holistic models assume that idioms form a discrete category of phrases that are reliably differentiated from non-idioms.

Swinney & Cutler's (1979) results supported this prediction - idioms demonstrated a facilitation effect such that meaningfulness judgements were faster for idioms than non-idioms. These findings also perpetuate holistic assumptions regarding an

idiom-specific processing mechanism. Holistic models do not propose special treatment of non-idiomatic collocations, which may or may not be holistically represented. Thus, idiomatic meaning retrieval is not considered to be a function of formulaicity, familiarity, or predictability. Instead, it is a processing requirement specific to idioms due to their noncompositionality. This difference highlights holistic reliance on the assumption that idioms form a cleanly delineated, discrete class. Thus, the idiom facilitation effect demonstrates the ability of native speakers to quickly recognize idioms and retrieve their holistically represented phrasal meaning.

At the heart of the assumption of holistic models, that idioms form a discrete, recognizable class, is the definition of an idiom as a noncompositional phrase. While this definition provides a clean, if simplified, description, idioms cannot be so easily boxed in. Nunberg et al. (1994) warned of the dangers of a simplified definition of idioms, noting that "idioms occupy a region in a multidimensional lexical space, characterized by a number of distinct properties" (Nunberg et al. 1994:492). While formulaicity and noncompositionality are shared by all members of the class idiom, they are not, on their own, sufficient to set idioms apart from other types of language. It is this insufficiency that led to the creation of individual analysis models, which consider similarities between idioms and other types of language as well as variation between phrases within the class idiom.

2.2. Individual analysis models

Recognizing the danger of reliance on a simplified definition of an idiom with respect to how we think about and understand these phrases, individual analysis models reject the idea of an idiom-specific mode of comprehension in which all idioms are handled in the same manner. Instead, they note that noncompositionality is not unique to idioms and argue that this property cannot

serve as a sufficient condition to describe the class idiom. This argument is reflected in individual analysis comprehension models: since it was noncompositionality, and the associated assumption that it delineates idioms as a unique class requiring specialized lexical representation, that necessitated direct retrieval via an idiom-specific comprehension mechanism, they further argue that language-general comprehension models can explain how we understand these phrases (Gibbs 1980, Gibbs & Nayak 1989, Gibbs et al. 1989, Caillies & Butcher 2007, Vilkaite 2016, Carrol & Conklin 2019).

There are three major models that I classify as individual analysis models, the idiom decomposition hypothesis (Gibbs & Nayak 1989, the conceptual metaphor hypothesis (Gibbs et al. 1997), and the hybrid model (Titone & Connine 1994b, 1999; Libben & Titone 2008; Titone & Libben 2014; Titone et al. 2019). Individual analysis models are united by their belief that idioms are not mentally represented as "words with spaces" (Gibbs & Nayak 1989, Gibbs et al. 1989). During comprehension, all words in an idiom are individually analyzed and each contributes to meaning activation (Gibbs & Nayak 1989; Gibbs et al. 1989; Titone & Connine 1994b, 1999; Gibbs et al. 1997, Caillies & Butcher 2007; Libben & Titone 2008, Titone & Libben 2014, Titone et al. 2019). Because individual analysis models do not assume that idioms must be holistically represented and stipulated as idiomatic or that they are easily differentiated from non-idioms by native speakers, they posit that it is not a phrase's status as a member of the class idiom that impacts how we think about and comprehend these phrases. Instead, they consider similarities and differences between idioms and non-idioms and propose that a number of properties, associated with, but not unique to, idioms impact our conceptualization of these phrases.

Individual analysis models reject the idea of a cleanly delineated class idiom based on noncompositionality for two reasons. First, they note that idioms are not the only kind of noncompositional phrase. Second, they recognize the richness of idiomaticity, noting that a number of additional properties are necessary to characterize idiomaticity.

Idioms are only one type of noncompositional language. For example, the meaning of a sarcastic or ironic utterance differs from the meaning conveyed by the individual parts. Less obvious examples come from phrases such as *purple car, pierced earring, or stuffed animal,* which few native speakers would argue are literal, despite the fact that their meaning is not entirely compositional. We understand *purple car* through its intersective meaning, viz., purple things that are cars. But, *purple* refers to only a portion of the *car*, namely the exterior. Because some of its conventionalized meaning comes from the construction itself, *purple car* is not entirely compositional since not all of the semantic meaning can be attributed to either the modifier or the head noun. *Pierced earring* highlights the problematic nature of assuming a natural class non-idiom, as such a classification is predicated on the assumption that the meaning of non-idiomatic phrases is derived by adding together the meaning of the constituent parts. However, this is not the case for the non-idioms *pierced earring* and *stuffed animal*. Similarly, the meaning of the phrase *pierced earring* is not "an earring that has a hole in it" and the meaning of the kind of *stuffed animal* that is sold in toy stores is not "a member of the kingdom Animalia that is stuffed". Thus, the fact that noncompositionality is not unique to idioms renders it insufficient to adequately define the class idiom. Because of these similarities, the line between phrase types is not always clear, as illustrated by the fact that differentiating idioms from non-idiomatic collocations has a controversial history (Kamp & Partee 1995).

Individual analysis models posit that a number of properties, not just noncompositionality, are required to truly define the class idiom. Research conducted within the individual analysis framework has considered the role of formulaicity, distributed idiomatic meaning, partial literality, dual meaning, familiarity, flexibility, figuration, transparency, affect, and contextual appropriateness (e.g., informality) (Schweigert 1985, Schweigert & Moates 1988, Gibbs & Nayak 1989, Gibbs et al. 1989, Schweigert & Cronk 1992, Cronk & Schweigert 1993, Titone & Connine 1994b, Gibbs et al. 1997, Titone & Connine 1999, Nippold & Taylor 2002, Papagno 2003, Caillies & Butcher 2007, Libben & Titone 2008, Fanari et al. 2010, Titone & Libben 2014, Bulkes & Tanner 2017, Geeraert et al. 2017, de Silva 2018, Kedzierska 2018, Carroll & Conklin 2019, Titone et al. 2019, Vulchanova et al. 2019, Chakrabarty at al. 2023). Findings indicate that formulaicity, distributed idiomatic meaning[9], partial literality, and dual meaning impact idiom conceptualization and comprehension, suggesting that they are the most likely to be reflective of the cognitive construct of idiomaticity.

However, no work to date has established a direct relationship between these idiom-related properties and idiomaticity that would help determine which properties are defining characteristics of idioms. Further, while individual analysis models share the assumption that multiple properties may impact idiom conceptualization and comprehension, they differ as to which properties are viewed as significant, leading to theory divergence regarding the role played by a given property. Although each

[9] Flexibility and figuration have been shown to have a strong correlation with distributed idiomatic meaning. However, Gibbs & Nayak 1989 and Gibbs et al. 1989 showed that flexibility and figuration do not directly impact online comprehension unless there is a processing error during the first attempt to comprehend a phrase. Instead, their impact is limited to metalinguistic tasks requiring conscious analysis. This suggests that any role they play occurs after initial processing (cf. Gibbs & Nayak 1989, Libben & Titone 1999, Carroll & Conklin 2019). Thus, while they may impact idiom conceptualization, their impact may be secondary to, or even reflective of, distributed idiomatic meaning. This question deserves further investigation but is outside the scope of this work. Crucially, Gibbs & Nayak (1989) showed that, by controlling for distributed idiomatic meaning, one neutralizes any impact of flexibility or figuration. Thus, this work considers distributed idiomatic meaning only.

property may individually impact comprehension in some manner, little is known about the relationships between the idiom-related properties and idiomaticity or even between properties themselves. Instead, a subset of properties is adopted by a given research group while others are ignored. Despite the general acceptance that multiple factors impact comprehension, this practice reduces the dimensionality of idiomaticity by adopting the assumption that the chosen properties are sufficient to investigate "idiomaticity". Thus, the second goal of this work is to directly investigate the relationship between idiomaticity and the four properties most likely to be among its defining characteristics.

The following subsections introduce the idiom properties that are most likely to be defining characteristics of idiomaticity, that is, those most likely to reflect the cognitive construct of idiomaticity and account for its continuous nature (cf. Titone & Connine 1994b, Libben & Titone 2008, 2014 Wulff 2008).

2.2.1. Formulaicity

The insufficiency of noncompositionality to differentiate idioms from non-idioms led Gibbs (1980) to argue for a continuum-based approach. In this approach, idioms are not "special", meaning that there is no idiom-specific comprehension strategy. Instead, idioms fit within the language-general comprehension framework. As such, a number of properties have been shown to impact comprehension. These properties may also be defining characteristics of idiomaticity.

Phrasal formulaicity is the first idiom property that, while not unique to idioms, impacts comprehension and is thought to impact their conceptualization. All idioms are, at least to some degree, formulaic (cf. Wray & Perkins 2000). However, idioms are only one subtype within the very large class of formulaic phrases. Other members of this class include verb particle constructions (ex. *turn on, light up, log*

on), non-idiomatic collocations (ex. *human rights, jam session, government spending*), and binomials (ex. *salt & pepper, bride & groom, aunt & uncle*). Formulaic phrases often have a conventionalized meaning[10]. For example, the non-idiomatic collocation *center divider* is a compositional formulaic phrase since the meaning of *center divider* is the sum of the literal senses of "center" + "divider" (Nunberg et al. 1994). Like idioms, the formulaic nature of non-idiomatic collocations leads to narrowed, conventionalized meaning. For example, the compositional combination of center+divider should include all items that divide something in the middle. However, most native speakers of American English immediately think of a highway median when they hear *center divider*, rather than perhaps a barrier in the middle of an ice rink, because it is the most conventional meaning.

Formulaicity is central to the question of categories defined by necessary and sufficient conditions versus a prototype structure[11] (Rosch 1973, Rosch & Mervis 1975, Geeraerts 1989, Kamp & Partee 1995, Wulff 2008, Pentilla 2011, Watson 2019, Rosch & Lloyd 2024). As previously discussed, the assumption of a cleanly delineated class idiom finds indirect support in the idiom facilitation effect. This interpretation attributes facilitation to noncompositionality. However, Gibbs (1980) challenged this interpretation, suggesting that facilitation could be due to something more general such as phrasal formulaicity. When particular words appear together more often than would be expected by chance, they become

[10] Conventionalized meaning refers to the shared interpretation of a widely used phrase within a community of practice. This meaning is not necessarily transparent or directly derivable from the meanings of the individual words of the phrase. Because the meaning of an idiom is not the sum of the parts, all idioms have conventionalized meaning (Nunberg et al. 1994, Wulff 2008). However, conventionalized meaning is not limited to idioms but can be seen with all types of phrases (e.g., literal, metaphoric, etc.) when the meaning is conventional or agreed upon by speakers.

[11] See section 5.4 for a discussion of the applicability of prototype-based approach as compared to a radial categories approach to this work and why a prototype-based approach was ultimately selected.

interdependent, or "internally sticky" and this "stickiness" allows for predictive cues to come from within the phrase itself, in addition to those that come from context.

Though initially met with harsh criticism, Gibbs' (1980) alternative explanation highlighted an important methodological consideration for experiments demonstrating the idiom facilitation effect. As its name implies, the idiom facilitation effect is thought to reflect a processing advantage unique to idioms, which arises when idioms are compared to non-idioms[12]. In experiments, it is standard practice to choose non-idioms with an orthographic form that is as similar to an idiom as possible. Ideally, this is accomplished by choosing idioms with plausible literal and idiomatic meaning then comparing reaction times for both phrases. Because not all idioms have a plausible idiomatic and non-idiomatic interpretation, an alternative approach is to change only one word in an idiomatic phrase to another, unrelated word of similar length, thereby blocking an idiomatic interpretation. While the reasoning behind this methodology is sound, it is problematic in practice. In order to meet these experimental control requirements, the non-idiomatic phrases that idioms are compared to differ significantly in their degree of formulaicity (see 1a. & 1b., from Cacciari & Tabossi 1988:682) and often require one to accept an unusual interpretation of a normally idiomatic phrase (see 2a. & 2b., from Geeraert et al. 2017:81). Both marginally acceptable meaning and differing degrees of formulaicity independently impact comprehension speed, regardless of whether comparison phrases are idiomatic or not.

1a. Once again, the boy <u>landed on his feet</u>. (Idiom)

1b. Once again, the boy <u>landed on his shoes</u>. (Non-idiomatic comparison)

[12] Referred to in the literature as "literal counterparts", a term equally problematic as many "literal counterparts" used in experiments are abstract, not actually literal.

2a. My friend is a notorious gossip and heard through the grapevine that they had broken up long before it was common knowledge. (Idiom)

2b. My friend is an avid gardener and heard through the grapevine that the plant needed water or it would die. (Non-idiomatic comparison)

More recently, Carrol & Conklin (2019) investigated the role of formulaicity in three types of formulaic language: idioms, non-idiomatic collocations, and binomials (e.g., *salt and pepper*). They demonstrated that facilitation is not limited to idioms but is also seen with other types of formulaic language, such as non-idiomatic collocations. These findings raise questions as to whether meeting the criteria for inclusion in an idiom dictionary is the most significant factor in the comprehension of idiomatic phrases or whether something more general, such as formulaicity, is responsible. Because all types of formulaic phrases are highly frequent and therefore well-learned, a speaker may rapidly map a specific construction to its meaning. When facilitation is seen with all types of formulaic language other than idioms, it is attributed to formulaicity. Only when facilitation is seen with phrases labeled as "idioms" is it attributed to something else; namely idiomaticity. Thus, Carrol & Conklin's (2019) findings challenge the holistic interpretation of the idiom facilitation effect, in which facilitation is reflective of a phrase's status as an idiom. Instead, they suggest that formulaicity, not idiomaticity, may account for a significant portion of this effect. Additionally, their findings challenge the necessity of an idiom-specific comprehension strategy. According to statistically based, language-general processing theories, there is no reason to believe that the idiom facilitation effect should be attributed to idiomaticity (Rammel et al. 2017, Vespignani 2020). Because predictability leads to processing ease, the more formulaic a phrase is, the more predictable it is, regardless of its status as idiomatic or non-idiomatic. Considering the central role played by prediction in language comprehension (and cognitive processes more generally), such an account seems

plausible – a predictable phrase is less metabolically expensive, a consideration that far exceeds the status of a phrase as idiomatic (Gazzaniga et al. 2014). Thus, the role of phrasal formulaicity cannot be ignored.

Despite the questions raised by Carrol and Conklin (2019), their findings alone cannot definitively disprove an alternative (holistic) account in which something other than noncompositionality is responsible for findings of faster comprehension of idioms versus non-idioms. In their experiments, Carrol & Conklin (2019) showed that formulaicity is a predictor of reaction time speeds for idioms, non-idiomatic collocations, and binomials. However, they did not directly compare idioms to non-idiomatic collocations or binomials because they did not identify a way to calculate formulaicity across phrase types and lengths[13]. Because idioms were not directly compared to non-idiomatic collocations or binomials and because prior existing work demonstrating the idiom facilitation effect did not control for formulaicity, attribution of idiom facilitation to idiomaticity cannot be dismissed. Unfortunately, while attribution of idiom facilitation to formulaicity has gained traction, no work to date has directly compared idioms with equally formulaic non-idioms.

2.2.2. Defining semantic characteristics

In addition to phrasal formulaicity, proponents of an individual analysis approach have identified at least three semantic properties of idioms that may characterize

[13] Phrases used by Carrol & Conklin (2019) differed in length between phrase types. All test idioms were two words in length and pointwise mutual information (PMI) was used to calculate phrasal formulaicity (for a discussion of PMI, see section 3.1.1). However, PMI cannot be used for phrases consisting of more than two words, such as binomials, without modifications to the equation (i.e., multinomial mutual information (MMI), see Van de Cruys (2011) for a discussion of in-development MMI equations as this metric has not yet been optimized and validated for use in psycholinguistic work). By virtue of the experimental design, Carrol & Conklin (2019) were not able to directly compare phrases across types. Instead, they chose to indirectly compare observed facilitation between phrase types, which allowed them to access other features of idioms, binomials, and non-idiomatic collocations in addition to formulaicity.

idiomaticity, thereby differentiating phrases within this class. These properties are distributed idiomatic meaning (previously referred to by a variety of names, including "compositionality", "decompositionality", and "global decompositionality"[14], partial literality and dual meaning (Gibbs & Nayak 1989; Gibbs et al. 1989; Titone & Connine 1994b,1999; Caillies & Butcher 2007, Libben & Titone 2008; Tabossi et al. 2008, Nordmann et al. 2013, Titone & Libben 2014, Bulks & Tanner 2016, Titone et al. 2019). These properties impact native speaker intuitions and comprehension of idioms, highlighting their import and further calling into question the assumption that all idioms are analyzed and mentally represented in the same manner (Gibbs & Nayak 1989; Gibbs et al. 1989; Titone & Connine 1994b,1999; Gibbs et al. 1997; Caillies & Butcher 2007, Libben & Titone 2008; Titone & Libben 2014). The fact that these properties impact conceptualization (evinced via shared native speaker intuitions) and comprehension (evinced via differences in comprehension speeds) makes them likely candidates to at least partially define the class idiom.

As previously mentioned, ***distributed idiomatic meaning*** refers to whether individual words in an idiomatic phrase contribute meaning or not. For example, *spill the beans* ("to unintentionally, prematurely, or indiscreetly reveal secret or privileged information" (oed.com)) has distributed idiomatic meaning because *spill* contributes the meaning of "to reveal" and *the beans* contributes the meaning of "information". However, *kick the bucket* ("to die") does not have distributed

[14] In the psycholinguistic literature, "decompositionality" is differently operationalized, alternatively referring to distributed idiomatic meaning, dual meaning, or partial literality. Additionally, some further subdivide decompositional idioms based on whether the phrase is figurative or not. These studies use the term "global decompositionality" to refer to the general category of decompositional idioms, labeling figurative decompositional idioms "abnormally decompositional" and non-figurative decompositional idioms "normally decompositional". To avoid confusion, "decompositionality" will not be used. Instead, "distributed idiomatic meaning" will be used to refer to what Nunberg et al. (1977, Wasow et al. 1984) labeled "compositionality", and later refined to differentiate between "idiomatically combining elements" (previously "compositional idioms") and idiomatic phrases (previously "noncompositional idioms").

idiomatic meaning because the meaning of "to die" is not contributed by either *kick* or *the bucket*.

Support for considering distributed idiomatic meaning as a potentially defining characteristic of idiomaticity comes from Gibbs & Nayak (1989), who showed that distributed idiomatic meaning impacts both how we think about idioms and how we comprehend them in real time. To determine whether distributed idiomatic meaning impacts idiom conceptualization, participants performed a metalinguistic judgement task, rating the degree to which individual words in an idiomatic phrase contribute meaning (e.g., *spill* ("reveal") *the beans* ("information")) or not (e.g., *kick the bucket* ("to die")). They found that native speakers are able to reliably categorize idioms based on distributed meaning, finding agreement rates of 75%. This indicates that native speakers share intuitions regarding distributed idiomatic meaning. Shared intuitions suggest that distributed idiomatic meaning is a common cognitive process and may influence phrasal representation, as the high agreement rates indicate that speakers consistently recognize and rely on the contribution of individual word meanings when interpreting idioms.This finding was surprising as native speakers were previously assumed to be entirely unaware of distributed idiomatic meaning, a belief that extended to comprehension as it was also assumed that distributed idiomatic meaning could not impact comprehension as native speakers were unaware of distributed idiomatic meaning at subconscious, preconscious, and conscious levels (Vulchanova et al. 2019, Milburn 2020). If this were true, any attempt to test these assumptions about distributed idiomatic meaning would be fatally flawed as there would be no way to create valid stimuli. In order to test distributed idiomatic meaning, researchers must construct a list of experimental stimuli in which phrases are categorized as having distributed idiomatic meaning or not having distributed idiomatic meaning. If native speakers

do not share intuitions, we cannot know whether an experiment truly addresses distributed idiomatic meaning or may instead be reflective of something else. Thus, the finding of Gibbs & Nayak (1989) was particularly significant as it validated further consideration of distributed idiomatic meaning[15].

After establishing that native speakers share intuitions about distributed meaning, Gibbs et al. (1989) took this conceptual-level work a step further, using the same phrases to demonstrate that distributed idiomatic meaning impacts not only conceptualization but also comprehension. A central tenet of the individual analysis approach is the belief that individual words in an idiom are analyzed. As such, analyzability should impact comprehension. Meaning associated with a known construction activates phrasal meaning for all idioms, regardless of whether a

[15] Additionally, while no work to date has considered the relationships between distributed idiomatic meaning, partial literality, and dual meaning, some have considered relationships between distributed idiomatic meaning and other features, such as idiom flexibility and figuration. For example, Gibbs & Nayak (1989) were interested in the relationship between properties that might explain why syntactic modification of some, but not all, idioms block an idiomatic interpretation of a phrase. To this end, they investigated the relationship between syntactic flexibility, figuration, and distributed idiomatic meaning. They found that syntactic flexibility is determined by shared intuitions between speakers regarding distributed idiomatic meaning. Specifically, they found that a phrase is significantly more likely to be flexible when at least one word contributes meaning similar to the meaning it carries outside of the phrase (e.g., *pop the question, lay down the law, foot the bill*) than when the distributed meaning is figurative (e.g., *promise the moon* "offer more than one can give", *spill the beans* "reveal information", *crack the whip* "to use authority", *line one's pockets* "to embezzle funds" (Gibbs & Nayak 1989:133). From this, they concluded that the ability to assign meaning to individual words in an idiomatic phrase explains native speaker intuitions regarding flexibility and hypothesized that distributed idiomatic meaning may be able to account for any impact of flexibility or figuration during comprehension. In a post-hoc analysis, they reanalyzed response times for phrases with distributed idiomatic meaning to determine whether flexibility or figuration played a significant role in comprehension. Despite finding a strong positive correlation at the conceptual level between distributed idiomatic meaning and flexibility when the meaning of a phrase is transparently related to its literal meaning and a strong negative correlation between distributed idiomatic meaning and flexibility when the distributed meaning of a phrase is figurative, flexibility and figuration did not significantly impact comprehension. Instead, they found that the analyzability of individual words in a phrase impacts comprehension, particularly when a phrase is semantically ill-formed (e.g, *pop the question, give the bounce*). Thus, there is no correlation between the ability to recognize a phrase as flexible or figurative and online comprehension. However, distributed idiomatic meaning does play a crucial role in how we think about and comprehend idioms. Additionally, the post-hoc analysis opened the door for further investigation into the assignment of meaning and the role it plays during comprehension.
*****Note that two of these three examples also demonstrate partial literality. Because Gibbs & Nayak (1989) did not investigate or control for partial literality, it is unclear what exactly was tested. Responses to phrases rated as having "literal distributed idiomatic meaning" may, in fact, be reflective of native speaker intuitions about distributed idiomatic meaning. However, it is likely that participants were instead picking up on the less cognitively taxing differentiation of determining whether a phrase has a word used literally (e.g. *pop the question*) or not (e.g. *blow your stack*) (Thibodeau et al. 2018). It is inconsistencies such as this that experiment 2 addresses.

phrase has distributed idiomatic meaning or not. However, when a phrase has distributed idiomatic meaning, individual words also contribute to the activation of the idiomatic representation. Thus, Gibbs et al. (1989) theorized that phrases with distributed idiomatic meaning should be comprehended more quickly than those with no distributed idiomatic meaning because activation of an idiomatic representation for phrases with distributed idiomatic meaning comes from two sources. The findings of Gibbs et al. (1989) supported this theory, showing that phrases reliably categorized as having distributed idiomatic meaning (Gibbs & Nayak 1989) were comprehended more quickly than those without distributed idiomatic meaning.

Partial literality refers to whether at least one word in an idiomatic phrase contributes the same meaning it has outside of the phrase. For example, *foot the bill*, is partially literal because *bill* refers to "an invoice" both within this phrase and outside of it while the meaning of *foot* in this phrase conflicts with its meaning outside of the phrase. However, *kick the bucket* is not partially literal because neither *kick* nor *the bucket* contributes the meaning that these words carry outside of the phrase. When an idiom is only two-words long, partial literality refers to whether exactly one word contributes a meaning inside of an idiomatic phrase that it has outside of the phrase (e.g., *pat answer*, which is a kind of *answer* but not one that can be understood by applying a meaning that *pat* has on its own outside of this phrase as a modifier).

Support for considering partial literality as a potentially defining characteristic of idiomaticity comes from studies demonstrating that native speakers are able to categorize phrases based on partial literality and finding that idiomatic phrases categorized as partially literal are comprehended more quickly than fully idiomatic phrases (cf. Gibbs et al. 1989, Titone & Connine 1999, Libben & Titone 2008).

According to Titone & Connine 1994b (see also 1999, 2014; Libben & Titone 2008), during comprehension, the literal meanings of individual words in an idiom are activated. When a phrase is partially literal the meaning of at least one word matches the activated literal meaning, providing two congruent sources of activation, which can facilitate meaning commitment and retrieval. By contrast, if neither word conveys a meaning it can have outside of the phrase, there is a conflict between the activated literal representation and each word in the idiomatic phrase. Conflicting meanings can delay the point at which a comprehender can commit to one meaning over another. Thus, if individual words in an idiom are analyzed individually, partial literality may significantly impact how we think about and understand idioms.

Proponents of the individual analysis approach posit that when an idiom is encountered, one must somehow determine that the combined literal meaning of individual words is not intended. The earlier this determination is made, the sooner idiomatic meaning can be pursued. While it is unclear whether literal analysis continues after a phrase has been recognized as idiomatic, experimental findings suggest that relatedness impacts how we think about and comprehend idioms (Titone & Connine 1994b, 1999). At the lexical level, overlapping meaning between a word in an idiomatic phrase and its meaning outside of the idiom is thought to provide two sources of activation – one literal and one idiomatic. For example, in the phrase *pop the question*, the meaning contributed by *question* is the same within the phrase and outside of it. If a literal analysis continues even after a phrase has been recognized as an idiom, then the meaning of *question* is activated via both the literal analysis and the idiomatic analysis (Cacciari & Tabossi 1988, Titone & Connine 1994a, 1994b, 1999).

The road to recognizing the import of partial literality began with Gibbs & Nayak (1989). In addition to considering the role of distributed idiomatic meaning, Gibbs & Nayak (1989) also tested for potential relationships between distributed idiomatic meaning, syntactic flexibility and figuration. At the conceptual level, they found a strong positive correlation between distributed idiomatic meaning and flexibility when the meaning of a phrase is transparently related to its literal meaning and a strong negative correlation between distributed idiomatic meaning and flexibility when the distributed meaning of a phrase is figurative. However, these findings did not hold for online processing; they found that figuration did not significantly impact comprehension. Instead, they found that the analyzability of individual words in a phrase impacts comprehension, finding faster comprehension of phrases in which one word contributed a meaning similar to the meaning conveyed outside of the phrase and concluding that the more apparent the meaning contributed by individual words in an idiomatic phrase is, the faster the phrase is processed. This post-hoc analysis opened the door for further investigation into the various ways in which meaning can be assigned and the role that assigned meaning plays during comprehension.

Excited by the potential role of within-phrase literality implied by Gibbs' post-hoc analysis, Titone & Connine (1994a, 1999) devised a series of experiments to directly test the role of partial literality in how we think about and comprehend idioms. Theorizing that overlapping semantic representation at the lexical level could result in strong activation and faster processing, Titone & Connine (1994a, 1999) separated phrases categorized as decompositional by Gibbs & Nayak (1989) into two groups: partially literal phrases (*pop* ("ask") *the question* ("question"), *foot* ("pay") *the bill* ("bill")) from those that were not partially literal (e.g., *lay down* ("to issue") *the law* ("strict orders"), *spill* ("reveal") *the beans* ("information")). They found

significant processing differences for partially literal phrases as compared to those in which no word was used literally (Titone & Connine 1999). They explained this finding by proposing that overlapping congruent meaning between a word in an idiomatic phrase and its meaning outside of the idiom provides two sources of activation – one literal and one idiomatic. Conversely, dual activation does not occur when a phrase is not partially literal. Instead, incongruent meaning at the lexical level impedes processing as both the literal and idiomatic meanings must be considered (Cacciari & Tabossi 1988, Titone & Connine 1999, Cailles & Butcher 2007, Libben & Titone 2008). In the case of *foot the bill*, if a literal analysis continues even after a phrase has been recognized as an idiom, then the meaning of *bill* would be activated during both the literal analysis and the idiomatic analysis, facilitating comprehension (Cacciari & Tabossi 1988; Titone & Connine 1994a, 1994b, 1999).

Dual meaning refers to whether a phrase has a plausible idiomatic and non-idiomatic interpretation or is plausible only when interpreted idiomatically. For example, *bad apple* has a dual meaning because it can be used idiomatically to mean something like "a bad or corrupt person in a group, typically one whose behavior is likely to have a detrimental influence on their associates" (oed.com) and non-idiomatically to refer to "an apple that is bad". However, *funny bone* can be used idiomatically only since there is no actual "bone that is funny". Consideration of dual meaning as a potentially defining characteristic of idiomaticity comes from studies demonstrating its impact on how we think about and comprehend idioms. Specifically, multiple, plausible representations of an idiomatic phrase (idiomatic and non-idiomatic versus non-idiomatic only) may impact meaning representation (Mueller & Gibbs 1987).

Dual meaning has been extensively studied, particularly by Titone and colleagues (Titone & Connine 1994a, 1994b, 1999, Libben & Titone 2008, 2011, Titone et al. 2014). Titone & Connine (1994a) began their investigation of dual meaning at the conceptual level. Using a ratings task, native speakers read 171 idiomatic phrases and rated them on one of 5 dimensions: their familiarity with a phrase, distributed idiomatic meaning, figuration, predictability, and dual meaning. The results of this experiment indicated that native speakers were not able to reliably categorize phrases based on dual meaning. However, they found a significant relationship between dual meaning and figuration. Specifically, they found a negative correlation between figuration and dual meaning such that phrases with dual meaning were significantly less likely to be deemed figurative. Interested in this finding which suggested that, individually, dual meaning might not reflect a common cognitive mechanism at work during this type of task but that also suggests dual meaning may be more readily accessible when a phrase is also figurative, this work was followed with a different type of experiment. In a new set of experiments collecting meaningfulness judgements, they found that distributed idiomatic meaning and dual meaning independently impact how we think about idioms (Titone & Connine 1999). From these experiments, Titone & Connine (1994a, 1999) concluded that dual meaning impacts how we think about idioms.

Seeking to better understand the role of dual meaning during online comprehension, Titone and colleagues (Titone & Connine 1994b, 1999, Libben & Titone 2008, Titone et al. 2014) built on these conceptual-level studies. Using a number of methodologies, including reaction time, reading time, and eye tracking, they posited that dual meaning would impact comprehension, theorizing that commitment to a representation would be easier when a phrase lacked dual meaning, (i.e., it had a plausible idiomatic interpretation only). In their

experimental work, they predicted that idioms with no dual meaning would be processed more quickly than those with dual literal and idiomatic meaning. Their results supported this theory, with findings demonstrating a processing advantage for phrases with no dual meaning over those with dual meaning (Titone & Connine 1994b, 1999, see also Libben & Titone 2008, Titone et al. 2014).

According to Titone & Connine (1994b), this finding indicates that plausible dual literal and idiomatic meaning requires activation and consideration of both representations. The degree of competition between the two kinds of meaning during online comprehension is impacted by transparency, familiarity, and predictability, which moderate the point at which commitment to one type of representation over another is likely to occur (Titone & Connine 1999, Titone & Libben 2014). Irrespective of the point of representational commitment, dual meaning impedes comprehension.

The impact of dual meaning differs from that of partial literality (Titone & Connine 1994a). When a phrase has no plausible literal interpretation, commitment to an idiomatic interpretation is fairly straightforward. However, when a phrase has a plausible literal and idiomatic meaning, both are activated (e.g., *bad apple* "a bad influence" and "a spoiled apple") and one must be chosen. While partial literality is thought to facilitate idiom comprehension, dual meaning delays the point at which one commits to an idiomatic interpretation (Titone & Connine 1999, Sprenger et al. 2007, Libben & Titone 2008, Titone et al. 2014).

If properties such as distributed idiomatic meaning, partial literality, and dual meaning impact how we think about and understand idioms can we still assume that all idioms will be recognized with equal reliability? Interestingly, native speakers' intuitions about certain properties of idioms are correlated with intuitions

about other properties. For example, phrases with no dual meaning are more likely to be viewed as having distributed idiomatic meaning than phrases that have dual idiomatic and non-idiomatic meaning (Titone & Connine 1994a). Considering the extent of variation within the class idiom as well as features of idioms that are shared with other types of language, such as formulaicity, it seems likely that there is no one-size-fits all mental treatment for idioms at any level of understanding – from how we think about and consciously analyze idioms, to their mental representation, to how their meaning is accessed. Instead, it may be more accurate to think of idioms as existing on a multidimensional continuum of prototypicality. I suggest that a phrase's placement on such a continuum may be based on certain combinations of features that make an idiom seem more or less prototypically idiomatic. At one end of the continuum are "prototypically idiomatic phrases" (e.g. *arm candy, sweet tooth*). This should correspond to phrases with no distributed idiomatic meaning, that are not partially literal, and that do not have a dual literal meaning, as lacking these properties make a phrase stand out as less analyzable and more dissimilar to a prototypical non-idiom. At the other end of the spectrum are "prototypically non-idiomatic phrases" (e.g., *weird places, fruit bowl*). This should correspond to normally compositional phrases, that are fully literal (both words contribute a meaning within the phrase that they convey outside of the phrase), and with no dual idiomatic meaning. In the middle are less prototypical members of either class. However, the straightforward approach of holistic models should not be dismissed without proper consideration.

2.3. Summary

The most basic point of contention between holistic and individual analysis approaches lies in the assumption that idioms form a discrete class that is reliably differentiated from non-idioms. According to holistic models, variation within the

class idiom and its potential impact on how we think about or comprehend idioms is not considered. During comprehension, each word is individually analyzed until the phrase is recognized as idiomatic, activating the idiom-specific processing strategy. Because a phrase must be recognized as an idiom in order for this strategy to be activated, these phrases must be stipulated as idiomatic in the mental lexicon. Thus, there must be a natural class idiom, the members of which are quickly and reliably recognizable to native speakers.

According to individual analysis models, variation within the class idiom is important. Because individual analysis models reject the idea of an idiom-specific processing mechanism, they posit that it is not a phrase's status as a member of the class idiom that impacts how we think about and comprehend these phrases. Instead, they believe that idioms fit within a language-general comprehension framework. As with other types of language, formulaicity, distributed idiomatic meaning, partial literality, and dual meaning are believed to impact how we think about and comprehend idioms as a lexical-level analysis is conducted. However, methodological inconsistencies have prevented individual analysis models from widespread acceptance. Specifically, a better understanding of the role played by formulaicity in how we think about and comprehend idioms versus non-idioms is crucial. Additionally, while idiom properties such as distributed idiomatic meaning, partial literality, and dual meaning have been shown to individually impact idiom conceptualization and comprehension, little is known about the relationships between these properties.

Until certain knowledge gaps are addressed, we cannot advance our understanding of the class idiom and more broadly, the processes involved in idiom conceptualization and comprehension. To this end, the present work asks two questions. The first research question (**RQ1**) seeks to determine whether there is a

categorical distinction between idioms and non-idioms. This is addressed in two parts. **RQ1a** questions shared intuitions of perceived idiomaticity, asking: **are native speakers able to differentiate between dictionary idioms and non-idiomatic collocations? RQ1b** builds on RQ1a to further refine our understanding of idiomaticity, asking: **do dictionary idioms receive higher ratings of idiomaticity from native speakers than equally formulaic non-idiomatic collocations?** In the following sections, experiment 1 addresses RQ1, focusing on testing the discreteness of the class idiom. By determining whether native speakers reliably differentiate idioms from equally formulaic non-idiomatic collocations, this investigation furthers our understanding of idiom conceptualization and addresses a critical knowledge gap by assessing the validity of a natural class idiom.

I predict that, overall, idioms, that are conventionally classified as such based on their inclusion in idiom dictionaries, will not be reliably differentiated from equally formulaic non-idiomatic collocations. This finding would be expected if the distinction between idioms and non-idiomatic collocations is not categoric. That is, I do not anticipate participants will be in high agreement that all dictionary idioms are idiomatic and all equally formulaic non-idiomatic collocations that are excluded from idiom dictionaries are non-idiomatic. Instead, judgements will be graded. A finding in which idioms are not differentiated categorically would necessitate further investigation as it may be that judgments of idiomaticity are correlated with multiple properties of idioms.

The second research question (RQ2) builds on current work demonstrating that, individually, specific properties of idioms impact how we think about and understand these phrases by considering the role played by each property. **RQ2** asks**: are formulaicity, distributed idiomatic meaning, partial literality,**

and dual meaning able to predict which phrases will be reliably recognized as idiomatic? This experiment focuses on testing whether and to what degree these idiom properties may be individually or collectively reflective of the mental construct of idiomaticity. Not only does it establish a foundation for future comprehension work, it provides a different approach to the mental category idiom, evaluating whether a prototype-based approach may be more appropriate than a natural-class approach. In line with prior individual analysis work, I predict a correlation between certain combinations of idiom properties and phrase type recognition ratings (Gibbs & Nayak 1989; Gibbs et al. 1989; Titone & Connine 1994a, 1999; Libben & Titone 2008; Tabossi et al. 2008; Nordmann et al. 2013, see also section 2.2.2). Such a finding would indicate scaler judgments, supporting a prototype-based approach.

CHAPTER III

EXPERIMENT 1

Experiment 1 addresses RQ1a and 1b, investigating the impact of idiomaticity on how we think about multiword phrases, seeking to determine whether native speakers share intuitions about idiomaticity. Specifically, RQ1a asks: are native speakers able to differentiate between idioms and non-idiomatic collocations? Building on this, RQ1b asks: do dictionary idioms receive higher ratings of idiomaticity from native speakers than equally formulaic non-idiomatic collocations? This investigation was accomplished by assessing the ability of native speakers to reliably differentiate idioms (e.g., *blank slate, sweet tooth*) from equally formulaic non-idiomatic collocations (e.g., *controlled chaos, ice cube*). To perform this assessment, participants completed a task in which they read a number of naturally-occurring sentences extracted from a 6-million-word, self-created corpus. Sentences ended in either an idiom or an equally formulaic non-idiom, and participants indicated whether the sentence-final phrase was an idiom or not.

If idioms comprise a cleanly delineated class, then native speakers would share intuitions regarding which phrases are members of this category and which are not. A finding in which idioms are reliably differentiated from non-idioms would indicate that idioms form a discrete class, indirectly supporting an account in which idiomaticity is synonymous with nondecompositionality. A finding in which idioms are not reliably differentiated from non-idioms would indicate that there is no clear bifurcation between idioms and non-idioms and would suggest that the mental

construct of idiomaticity is more complex than can be accounted for by a single variable.

RQ1a was investigated using Fleiss' kappa, a measure of inter-rater agreement reliability between three or more raters. I predict that a test of inter-rater reliability will show that native speakers are not able to reliably, categorically differentiate idioms from equally formulaic non-idiomatic collocations (**H1**). Thus, H1 predicts that, at least at the conceptual level, there is no clear bifurcation between idioms and non-idiomatic collocations such that there is a class of phrases, all of which are equally idiomatic, and another class of phrases, all of which are equally non-idiomatic. Instead, I propose that idiomaticity is a gradated construct. Within linguistics, differentiating idioms from non-idiomatic collocations has a controversial history (Kamp & Partee 1995). This prediction is consistent with linguistic work demonstrating that the line between idioms and other kinds of lexicalized phrases is not always clear (cf. Kamp & Partee 1995, Nenonen 2007, Wary 2013).

Crucially, such a finding would not indicate that there is no such thing as idiomaticity. If idioms exist on a continuum, significantly different ratings of idiomaticity to items within a condition (idioms versus non-idiomatic collocations) would be expected. RQ1a addresses the question of whether idiomaticity is a discrete, binary class only. It is unable to add weight to the question of prototypicality effects, or whether the class idiom is better accounted for as a continuous variable with certain members perceived as more idiomatic than others and how it is that we might account for differences in perceived levels of idiomaticity. When highly representative members of the class idiom are grouped with phrases exhibiting fewer representative properties, a homogenized analysis

would show the averaged effects. Unreliable recognition of the "averaged idiom[16]" would indicate only that the single, binary variable of idiomaticity is unable to account for differentiation, challenging the notion of a discrete category idiom. Instead, findings from such an analysis would indirectly support a prototype-based account in which only certain items are reliably recognized. In a prototype model, significantly different responses to items within a condition would be expected because items differ in the degree to which they align with features associated with a high degree of prototypicality. It is this eventuality that RQ1b considers, evaluating whether idiomaticity is continuous.

RQ1b will be investigated using a mixed effects logistic regression, which estimates the probability that a phrase will be rated as idiomatic. I predict (**H2**) that the model will reveal a significant effect of phrase type such that dictionary idioms are more likely to be rated as idiomatic compared to non-idiomatic phrases. However, this relationship will not be absolute. Ratings will show a high degree of variation between phrases and participants, with some dictionary idioms rated as less idiomatic than some non-idiomatic collocations and some non-idiomatic collocations rated as more idiomatic than some idioms. Such a finding would indicate that native speakers are, to an extent, aware of some sort of difference between idioms and non-idioms. It is simply that conceptual categorization of phrases by type is not shared between speakers, demonstrating the insufficiency of a general notion of idiomaticity as an explanatory variable.

Additionally, this analysis considered the role of formulaicity. As previously discussed, formulaicity refers to the probability that two words appear together intentionally rather than by chance simply because each has a high individual

[16] In this case, an "averaged idiom" refers to phrases expected to be highly representative, moderately representative, and marginally representative.

frequency. Formulaicity is not limited to idioms but is also seen with other types of non-idiomatic phrases. For example, the non-idiomatic phrases ice cube and catchy slogan are highly formulaic, as are the idioms sweet tooth and funny bone. By contrast, the phrases the boy and I went are not formulaic despite being comprised of some of the most frequent words in English. Idioms, as formulaic expressions, are known to enjoy a processing advantage due to their highly predictable nature, which is often reflected in faster comprehension (Gibbs 1980, Carrol & Conklin 2019). However, it is unclear as to whether formulaicity impacts idiom conceptualization. Within the class idiom, phrases vary as to their degree of formulaicity, with some phrases having a relatively low degree of formulaicity (e.g., real character) while others have a relatively high degree of formulaicity (e.g., sweet tooth). Some posit a positive correlation between formulaicity and perceived idiomaticity, such that the higher the degree of formulaicity, the higher the degree of perceived idiomaticity This would mean that formulaicity is a defining feature of idiomaticity. However, such a prediction assumes that conventionalization automatically leads to idiomaticity as it assumes that highly formulaic collocations become idiomatic due to conventionalization. While it is true that conventionalization narrows meaning, there must be a reason for a phrase to take on a noncompositional meaning. Motivations include recurrent, socially engaging themes and taboo or emotionally charged topics (Nunberg et al. 1994). Neutral phrases, or phrases that do not convey a personal or social stance, lack such motivation. This explains why phrases such as *ice cube* and *center divider*, while highly formulaic, do not seem particularly idiomatic and cautions against assuming a generalizable association between phrasal formulaicity and perceived idiomaticity. However, it should also be noted that, when sufficient motivation exists, formulaicity *is* associated with a higher degree of perceived idiomaticity. Thus, the relationship between formulaicity and idiomaticity is complex and may

not be captured via generalizations relying on straightforward relationships at the group (idiom or non-idiomatic collocation) level.

If there is a clear, categoric delineation between the class idiom and the class non-idiom, then the degree of phrasal formulaicity should not impact ratings. However, if idiomaticity is gradated, such that some phrases are seen as more idiomatic than others, then formulaicity may impact ratings. While formulaicity has been shown to impact comprehension, it is not expected to significantly impact analytical judgments at the class level (Bridges & van Lancker Sidtis 2013, Carrol & Conklin 2019, Eyigoz et al. 2020, Carrol 2023, Goldberg 2023). This is because, just as there is no clear relationship between familiarity and perceived idiomaticity, there is no clear relationship between the degree of formulaicity and perceived idiomaticity (cf. van Lanker 1987, Gibbs & O'Brian 1997, Wray & Perkins 2000, Wulff 2008, Snider & Arnon 2012, Carrol & Conklin 2019, Carrol 2023). For example, while *ice water* is highly formulaic, most native speakers would agree that it is not idiomatic. I predict (**H3**) that PMI will not significantly impact ratings. Such a finding would be in line with work demonstrating the complex relationship between idiomaticity and frequency-based measures, such as formulaicity, predictability, and familiarity. It would indicate that, on its own, formulaicity is not a predictor of perceived idiomaticity. Future work should consider the relationship between individual items and formulaicity to determine more fine-grained relationships.

3.1. Methodology

In line with prior work on idiom comprehension, a ratings task was used to determine whether idioms are reliably differentiated from equally formulaic non-idiomatic collocations. Participants read 144 naturally occurring sentences that ended with a phrase in bold. Participants rated the phrase in bold as "idiomatic" or "non-idiomatic", with the option of responding "unsure" if they were unable to make

a decision. Three types of phrases were included: idioms, non-idiomatic collocations, and fillers. All test items were 2-word phrases, had the form of either adjective + noun or noun+noun, and were balanced for orthographic length.

3.1.1. Materials

While RQs 1a and 1b were addressed by considering perceived idiomaticity and formulaicity, RQ2 incorporates the additional variables of distributed idiomatic meaning, partial literality, and dual meaning. To be able to model relationships between perceived idiomaticity and the investigated idiom properties, it was critical that all experiments include identical sentences and test phrases. Because the same stimuli were used in all experiments, stimuli needed to be carefully balanced so that the individual impact of a given variable could be investigated as could the relationships between variables. The following subsection explains stimuli creation in detail, beginning with idiom selection, moving to non-idioms, and concluding with the rationale behind filler creation. This is followed by an explanation of the phrase type recognition task, which was used in experiment 1 only.

The first step was to identify two-word idiomatic and non-idiomatic collocations with matching pointwise mutual information scores. Pointwise mutual information (PMI) is an information theoretic measure that quantifies the strength of association and co-occurrence between two words (Shannon 1948, Fano 1961, Church & Hanks 1990). Essentially, it compares the likelihood of word X and word Y appearing together intentionally (e.g., *bad apple*) with the likelihood of these words appearing together by chance due to high individual frequency (e.g., *boy that*, Church & Hanks 1990). Higher PMI values indicate a stronger association, suggesting that one word is more predictable given the other and indicating that words are more likely to be encountered as a fixed expression. Crucially, PMI is optimized for two-word phrases. While some have used it to quantify association

strength between 3 or more linguistic items, such use has not been rigorously tested (Lin 1999, Ellis et al. 2008). With respect to idioms, quantification of co-occurrence and association strength at a level equal to that of two-word phrases would necessitate equation modifications prior to implementation. Additionally, without more extensive modifications, PMI scores between word lengths are not comparable. For example, while minor changes to the equations might allow for equally accurate quantification of two-word and three-word phrases, PMI scores for two-word phrases cannot be directly compared to an identical score for a three-word phrase. This is because PMI quantifies the association between two words by comparing their joint probability to the product of their individual probabilities. When extending PMI to multi-word phrases, such as three-word phrases, the calculations involve joint probabilities of three words occurring together, which are typically higher for three-word idioms than those for two-word phrases. This results in higher PMI values for three-word idiomatic phrases compared to two-word phrases, even if the actual association strength is similar. Consequently, a PMI score of 4 for a two-word phrase may indicate a moderate association, while the same score for a three-word phrase might suggest a stronger association. Therefore, PMI scores across different phrase lengths are not directly comparable without appropriate adjustments. For these reasons, only two-word phrases were included in this research.

A list of 550 two-word nominal idioms (ex. *book worm, silver lining, split second, puff piece*) was created by consulting idiom dictionaries (Longman 1998, idioms.thefreedictionary.com) and by recording idioms identified in everyday conversations using native speaker intuitions and verified by consulting idiom dictionaries. Next, the PMI value of each idiom was calculated, using a six-million-

word corpus of informal language[17], created by scraping blog posts, review websites, movie scripts, and television transcripts. To identify potential non-idiomatic collocations, each idiom was matched with 50 two-word candidate phrases that had a PMI value identical to that of the idiom. Candidate phrases were narrowed down based on a number of criteria until only one PMI-matched non-idiomatic collocation (ex. *rough draft*) remained. For example, the ideal PMI-matched non-idiomatic collocation had to have the same structure as its idiomatic match (i.e. adjective+noun or noun+noun), could not include determiners (e.g., *the* or *a*) or pronouns (e.g., *its*, *your*, etc.), could not be idiomatic and needed to be of a similar length as its idiomatic match (for more details, see Appendices I and II).

Once paired, each idiom was analyzed along the three dimensions. Phrases were annotated as having distributed idiomatic meaning (e.g., *fresh mouth, spill the beans*) or not (e.g., *slam dunk, kick the bucket*), being partially literal (e.g., *acid test, foot the bill*) or not (e.g., *fresh mouth, slam dunk*), and as having plausible dual idiomatic and non-idiomatic meanings (e.g., *thin ice, green light*) or not (e.g., *funny bone, sweet tooth*). Next, phrases were classified based on their PMI values as having high PMI (PMI>6), medium PMI (5.9<PMI>3), or low PMI (PMI<2.9) (value ranges adopted from Hunston 2002, Durrent & Doherty 2010, Carroll & Conklin

[17] This corpus was created specifically for this project. It was initially created by Zak Boston, a CLASICS student who graduated in 2021. All blog posts and reviews were scraped by Zak, using standard NLP methods. In addition, Zak identified a large portion of the movie scripts and wrote the initial code to calculate PMI. Help was rendered as part of his CLASICS capstone project, with permission granted by our respective advisors. While Zak ultimately decided to accept a different capstone offer, I am extremely grateful for his amazing help! To his corpus, I added transcripts from additional television shows and movies. I also rewrote the python script used to calculate PMI to improve calculation accuracy. In this experiment, all phrases are paired based on the PMI scores calculated by my program based on phrases found in my corpus. However, to validate these numbers, my PMI values were compared to COCA (https://www.english-corpora.org/coca/, Davies 2008) and Sketch Engine (https://www.sketchengine.eu//, Kilgarriff et al. 2004, 2014). Sketch Engine values are available to anyone with an active Sketch Engine account, allowing for transparent replication. COCA values must be purchased. While the data cannot be shared, permission to share values can be granted by contacting Mark Davies (mark.davies@corpusdata.org; see also https://www.english-corpora.org/mutualInformation.asp for a comparison of PMI as calculated by various resources). Finally, my PMI program is hosted in CoLab and will be shared upon request.

2019). Finally, idiomatic phrases were embedded in the final position of a naturally occurring sentence.

While non-idiomatic collocations do not, by definition, have distributed idiomatic meaning and can not be partially literal[18] (since both words are used non-idiomatically), they can have dual meaning. All candidates were annotated as having plausible dual literal and idiomatic meaning (e.g., *thin ice*) or not (e.g., *human rights*). Next, non-idiomatic collocations were embedded in the final position of a naturally occurring sentence.

For the planned analyses, at least two tokens of each possible combination of the investigated properties were needed. For example, to control for potential influences of formulaicity, distributed semantics, partial literality, and dual meaning at once, phrases in the idiom condition needed to be balanced such that 24 of the 48 idiomatic phrases had distributed semantics and 24 did not, 24 of the idiomatic phrases were partially literal and 24 were not, and 24 of the idiomatic phrases had dual literal and idiomatic meaning while 24 did not. Additionally, 16 idioms needed to have high PMI, 16 had medium PMI, and 16 had low PMI (see Table 1 for the number of total stimuli by type and Appendix III for a visual representation of the factorial design). It was this requirement that necessitated the large number of phrases initially considered, as certain combinations of features are more common than others. Thus, the next step was to consider possible candidates for each combination (e.g., had distributed idiomatic meaning, was partially literal, had dual meaning, and had high PMI; did not have distributed idiomatic meaning, was partially literal, had dual meaning, and had high PMI; had distributed idiomatic

[18] Non-idiomatic collocations can not be partially literal under the adopted definition of partial literality used in this work. This definition is unrelated to degrees of concreteness and makes no claims about the various factors outside of idiomaticity that may impact "literality".

meaning, was not partially literal, had dual meaning, and had high PMI, etc.) and retain two for inclusion in this study, which was the minimum number of tokens needed to be able to perform the planned analyses.

A number of factors influenced final candidate selection, including phrasal familiarity, contextual support, individual word length and phrase length. For example, members of the Graduate Student Research Workshop[19] rated phrasal familiarity on a 5-point scale, where 1 indicated the phrase had never been heard and the meaning was unknown and 5 indicated that the phrase was often encountered and the rater was confident enough in its meaning to use it in contexts other than that in which they most often heard it. Phrases with a rating of less than 3 were removed. Additionally, length had to be considered. All phrases were embedded in naturally occurring sentences with context that clearly supported the intended interpretation. However, some phrases required more context to fully support the intended interpretation. The number of words in each carrier sentence was calculated and phrases differing in length by $\leq$1 standard deviation were reconsidered. If an equally supportive sentence closer in length to the average could be identified, the phrase was replaced by the better alternative, otherwise the phrase was removed altogether.

[19] The Graduate Student Research Workshop is an academic and professional development workshop series where students can request feedback on their work from their peers.

Condition	DIS	PL	DM	PMI
Idiom	24 +DIS 24 -DIS	24 +PL 24 -PL	24 +DM 24 -DM	8 High 8 Mid 8 Low
Non-idiom	48 -DIS	48 -PL	24 +DM 24 -DM	8 High 8 Mid 8 Low

Key:
DIS: distributed idiomatic meaning
PL: partial literality
DM: dual meaning
+: displays a property
-: does not display a property
High: PMI value greater than or equal to 6
Mid: PMI value of greater than or equal to 3-5.9
Low: PMI value less than or equal to 2.9

Table 1. Stimuli and experimental design. The idiom condition utilized a fully factorial design. The non-idiomatic collocation condition was balanced for dual meaning and PMI.

The final step was to create filler items. A common practice in the idiom comprehension literature is to compare idioms with non-idioms that differ by only one word (e.g., *land on your feet -> land on your shoes*). Filler phrases followed this format, substituting one word with a related word that blocked an idiomatic interpretation. In addition to being non-idiomatic, fillers had a PMI value of .9≤PMI, indicating that they do not meet the criteria to be considered a known collocation. 48 filler sentences were created, 24 of which were modified idioms (e.g., *movie screen,* from *silver screen*) and 24 of which were modified non-idiomatic collocations (e.g., *crude draft* from *rough draft*). Like the idioms and non-idiomatic collocations, fillers were embedded in a naturally occurring sentence, the average length of which matched the average length of the idiomatic and non-idiomatic collocation carrier sentences.

3.1.2. Phrase type recognition task procedure

A ratings task was created using Qualtrics. After providing consent to participate, a participant was taken to the screening and demographic questions. The screening questions asked for a participant's age; language experience, including their native language(s), and language(s) currently used for the majority of communication; whether they had participated in research related to idioms and non-idioms in the past year; and ability to see text on a computer screen (vision). These questions were asked for the following reasons. First, language use changes over the life span and this effect may be particularly pronounced for figurative language (Brysbaert et al. 2016, Sprenger et al. 2019). As such, only adults over the age of 18 but under 65 were eligible to participate. Second, this experiment relied on native language intuition. Languages learned after the age of 7 or a native language that is no longer the language of predominant use strongly impact responses (Athanasopoulos et al. 2015). As such, only native speakers of English who currently use English as their predominate form of communication were eligible to participate. Third, recent participation in idiom-related research, particularly in another ratings task included as part of this dissertation, could impact responses. Only those who had not participated in research related to idioms, figurative language, or collocations in the past year were eligible to participate. Finally, this experiment relied on visual acuity sufficient to read multiple sentences per screen for an extended period of time (~45-60 minutes). Only those who self-report as having normal or corrected-to-normal vision were eligible to participate.

The demographic questions asked a participant's gender, if they were right- or left-handed, if they had any underlying neurologic conditions, if they spoke any other language(s) in addition to their native language, and if they had any formal linguistics training. Answers to these demographic questions are necessary to rule

out the impact of variables known to impact language experiments but that are not directly studied or controlled for in this work. For example, I followed the common practice of collecting handedness responses due to a possible, but unexpected, confounding finding suggesting that those who are left-handed may be better at processing figurative language than those who are right-handed (Coulson & Van Petten 2002). Similarly, gender; neurological conditions including autism, dyslexia, and ADHD; language knowledge; and formal linguistics training have also been shown to impact linguistic studies and are expected to be taken into account prior to or during data analysis to avoid attribution of an effect to common individual differences (Cain et al. 2005, Siyanova-Chanturia 2011, Columbus et al. 2015, Cacciari et al. 2018, Vulchanova 2011, Vulchanova et al. 2015, Vulchanova et al. 2018). The impact of these factors was assessed for experiments 1 and 2 and were found to be insignificant. As such, no further discussion of screening or demographic information will be included.

After responding to these questions, qualified participants progressed to the task instructions, which explained that they would be asked to determine whether a number of phrases were idioms or non-idioms and provided a number of examples. This screen was followed by four practice sentences. Participants submitted their responses to the practice sentences and were shown the correct responses as well as an explanation as to why each phrase was classified as such. After completing the practice, participants advanced to the task, where they read each of the 144 sentences and rated the final phrase (which was marked in bold to draw their attention) as "idiomatic" or "non-idiomatic". Following feedback from an early pilot, an additional response option of "unsure" was added to reduce frustration. However, participants were asked to use the "unsure" response no more than 5 times throughout the entire survey. To ensure that participants were reading the

sentences, two attention checks were also included. Completion of the entire task took about 45 minutes. (See Appendix IV for an example of the task.)

3.1.3. Participants

In total, 115 participants completed this study. Participants were recruited from Prolific and CU Boulder. Prolific is an online community that matches psycholinguistic researchers with non-experts interested in participating in scientific work. In total, 40 participants were recruited from Prolific. Each Prolific participant received $12 as compensation for their time. 75 participants were recruited from CU Boulder using the linguistics department extra credit pool in the fall of 2022 and spring of 2023. All students were enrolled in LING 2000: Introduction to Linguistics. Students were given the option of receiving extra credit in their linguistics course or receiving a $12 e-gift card to Amazon or Starbucks.

3.2. Results and discussion

To determine whether there is a cleanly delineated class idiom reflected in shared native speaker intuitions about class membership, the data was analyzed using Fleiss' kappa and mixed effects logistic regression analysis. Fleiss' kappa, a measure of inter-rater reliability for 3 or more raters when chance is factored out (Fleiss 1971), was used to determine whether native speakers share intuitions about category boundaries by assessing whether idioms can be reliably differentiated from equally formulaic non-idiomatic collocations based on shared intuitions of category members (H1). A mixed effects logistic regression analysis was then employed to investigate whether variability in idiomaticity ratings can be explained by phrase type (idiom versus equally formulaic non-idiomatic collocation) and the degree of formulaicity (low PMI, medium PMI, and high PMI) to determine whether native speakers share a general awareness of differences between idioms and non-idioms, even if they do not share conceptual category boundaries (H2).

Differences in predicted probabilities were used to assess the relationship between idiom type and idiomaticity ratings, revealing nuanced differences in how native speakers rate dictionary idioms compared to non-idiomatic collocations. The role of formulaicity is considered in both analyses (H3).

3.2.1. Inter-rater reliability

To address RQ1a (H1), the data was analyzed using Fleiss' kappa. Fleiss' kappa is a statistical measure used to evaluate the reliability of agreement between three or more raters when assessing categorical ratings of a number of items. It is an extension of Cohen's kappa, which is used for situations involving no more than two raters. Fleiss' kappa was chosen over a more basic metric, such as average agreement, because it measures the agreement between raters beyond that which would be expected by chance alone. This is done using the following equation:

$$K = \frac{P_o - P_e}{1 - P_e}$$

Here, P_o is the observed proportion of agreement among raters, and P_e is the expected proportion of agreement by chance. P_o-P_e returns the degree of agreement above chance. 1-P_e is a normalization factor that corrects for the possibility of chance agreement. A kappa value of 1 indicates perfect agreement, while a value of 0 suggests that any agreement is purely by chance. Within the social sciences, a value of .75 has been established as the reliability cutoff for trustworthy data, with values over .75 accepted as highly reliable and values under .75 deemed unreliable (cf. Landis & Koch 1977, Shrout & Fleiss 1979; for further details on the kappa value interpretation, see scale in Figure 2).

3.2.1.1. Analysis

Of the 115 surveys collected, data from 106 participants was included in the analysis. Three tasks from Prolific users were excluded due to a technology malfunction and two were excluded due to failed attention checks. Four tasks completed by CU students were excluded due to failed attention checks.

Fleiss' kappa analysis was completed using DataTab (DataTab Team 2024). This analysis was structured such that agreement to all idioms and equally formulaic non-idiomatic collocations was assessed. Overall agreement was first assessed. This was followed by condition-specific (idiom versus non-idiom) agreement, where inter-rater agreement within each condition was assessed. While condition-specific agreement required grouping by phrase type, there was no presumption that responses would match that condition as Fleiss' kappa does not assess "correctness". For example, if all participants rated a phrase in the non-idiom condition as idiomatic, its kappa value would be k=1, reflecting perfect agreement. Therefore, evaluating agreement by condition does not assume that a phrase's inclusion in an idiom dictionary guarantees it will be universally recognized as an idiom.

3.2.1.2. Results and discussion

The results support the prediction that there is no discrete class idiom, finding that native speakers were not able to reliably differentiate idioms from equally formulaic non-idiomatic collocations (see Figure 2). The reliability cutoff for Fleiss' kappa is $k \geq .75$ (Landis and Kock 1977). Overall, inter-rater agreement for idioms and equally formulaic non-idiomatic collocations had a fair, but unreliable, agreement level of k=.25. The kappa value was significantly lower when considering responses to phrases within the dictionary idiom condition only and within the equally formulaic non-idiom condition only. Inter-rater agreement for idioms was k=.12 ($p<.001$), indicating that native speakers do not share strong intuitions about

members of the class idiom. Inter-rater agreement for non-idiomatic collocations was k=.12 (p<.001), suggesting that non-idiomatic status does not increase the consistency of responses. In both conditions, the observed unreliability was significant at p<.001. This indicates that the observed level of agreement was significantly different from chance. In other words, while ratings showed a "poor" level of agreement, native speakers do share some degree of intuition regarding members of the class idiom; the observed agreement was not due to random chance. Additionally, a standard error of 0 for both conditions indicates that there is little to no uncertainty in the kappa estimate (McHugh 2010, Jesussek & Jesussek 2024).

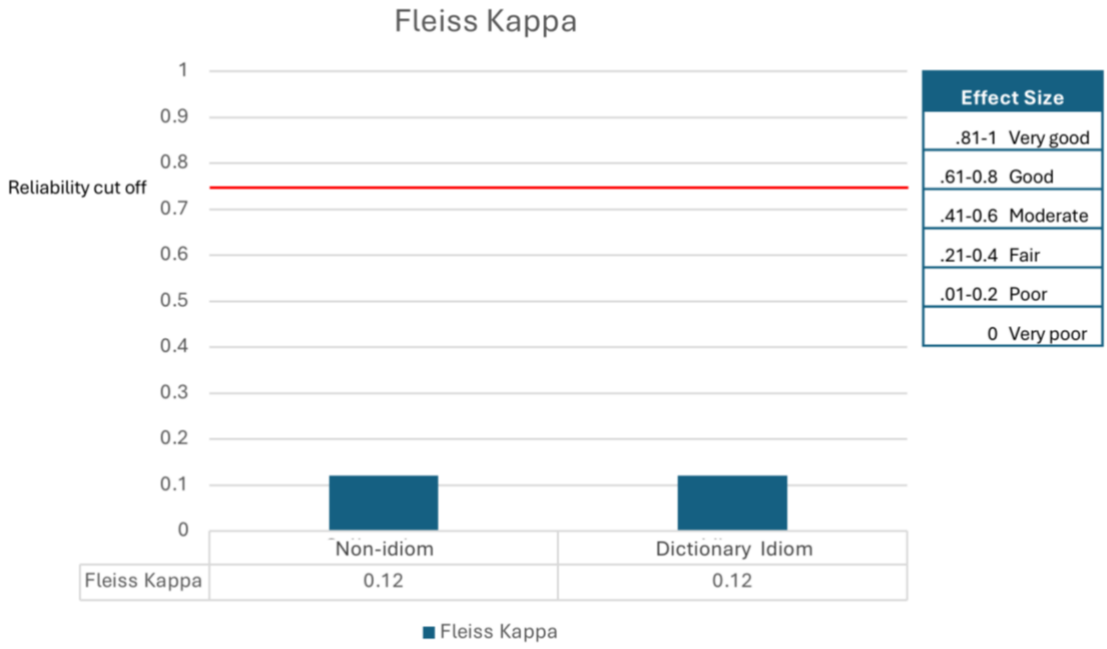


Figure 2. Fleiss' kappa values. Figure 2 shows Fleiss' kappa for ratings for phrases in the equally formulaic non-idiomatic collocation condition and in the dictionary idiom condition.

To be sure that ratings reflected perceived idiomaticity and were not reduced to a more straightforward judgement of phrasal predictability (cf. Thibodeau et al. 2018), the role of formulaicity was also considered as it could be that perceived idiomaticity was impacted by phrasal formulaicity. While no effect was anticipated, this analysis was critical as the role of formulaicity in online comprehension has been established (Carrol & Conklin 2019). Implementing PMI as a discrete variable, Fleiss' kappa was recalculated for each of the three PMI bands: low PMI phrases, medium PMI phrases and high PMI phrases (see Table 2). No effect of PMI was observed for any level of PMI.

Fleiss by PMI level						
PMI level	Condition	Fleiss Kappa	Standard Error	Lower 95% CI	Upper 95% CI	Fleiss' kappa p
Low	Non-idiomatic collocations	0.14	0	0.14	0.15	p<.001
	Idioms	0.19	0	0.18	0.19	p<.001
Medium	Non-idiomatic collocations	0.15	0	0.14	0.15	p<.001
	Idioms	0.11	0	0.11	0.12	p<.001
High	Non-idiomatic collocations	0.07	0	0.06	0.07	p<.001
	Idioms	0.07	0	0.06	0.08	p<.001

Table 2. Fleiss' kappa scores for idioms and equally formulaic non-idiomatic collocations based on PMI

Agreement for low PMI equally formulaic non-idiomatic collocations was k=.14. Agreement for low PMI idioms was k=.22. For medium PMI phrases, agreement for equally formulaic non-idiomatic collocations was k=.15 while agreement for idioms was k=.11. Agreement for high PMI equally formulaic non-idiomatic collocations

was k=.07 and agreement for high PMI idioms was k=.07. Differences between PMI bands were not significant (p=.09).

Overall, these results suggest that native speakers do not share intuitions about category boundaries between idioms and equally formulaic non-idiomatic collocations, regardless of the level of formulaicity. Further, they challenge a discrete account of idiomaticity. If there is a true categorical distinction between idioms and non-idiomatic collocations, then this should be reflected in a graph of the ratings with a clear delineation between the two groups. One would expect to see distinct clusters or peaks representing ratings for idiomatic phrases and equally formulaic non-idiomatic collocations, with little overlap between the two groups. This would indicate that participants consistently rated all idioms differently from all non-idioms, with little overlap between the two groups. Figure 3 presents a hypothetical graph depicting such a distinction. However, this is not what was observed in the actual distribution of responses. Instead, observed responses showed gradation, with ratings of a number of phrases falling between the two categories. This can be seen in Figure 4, which shows that certain idiomatic phrases were seen as more idiomatic than others and certain non-idiomatic phrases were seen as more non-idiomatic than others. This adds to the finding of unreliable inter-rater agreement, further indicating that there is no clearly delineated class idiom. Instead, it demonstrates the continuous, gradated nature of idiomaticity. Additionally, visual inspection shows that more dictionary idioms are rated as idiomatic than non-idiomatic collocations, indicating a general recognition of idiomaticity. From this, one must ask questions such as, "why are some items rated as idiomatic more reliably than others?" and, "is it that, overall, idioms are more likely to be rated as idiomatic than non-idiomatic collocations despite poor agreement at the item level?" Such questions motivated a second analysis, which

used a mixed-effects logistic regression to further evaluate the relationship between phrase type and idiomaticity.

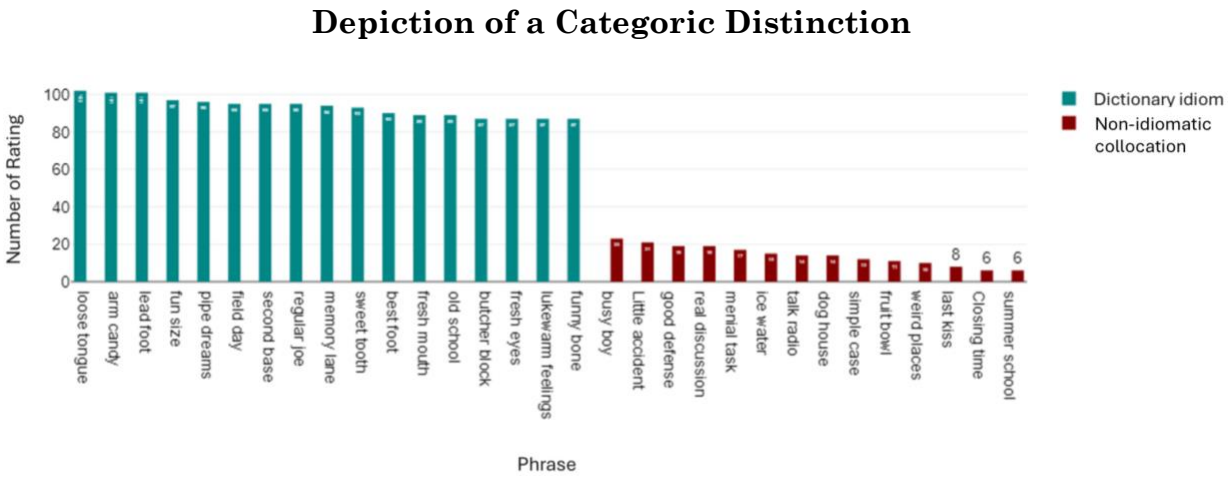


Figure 3. Hypothetical representation if responses were categorical. Figure 3 demonstrates the expected response distribution if there is a clearly delineated class idiom

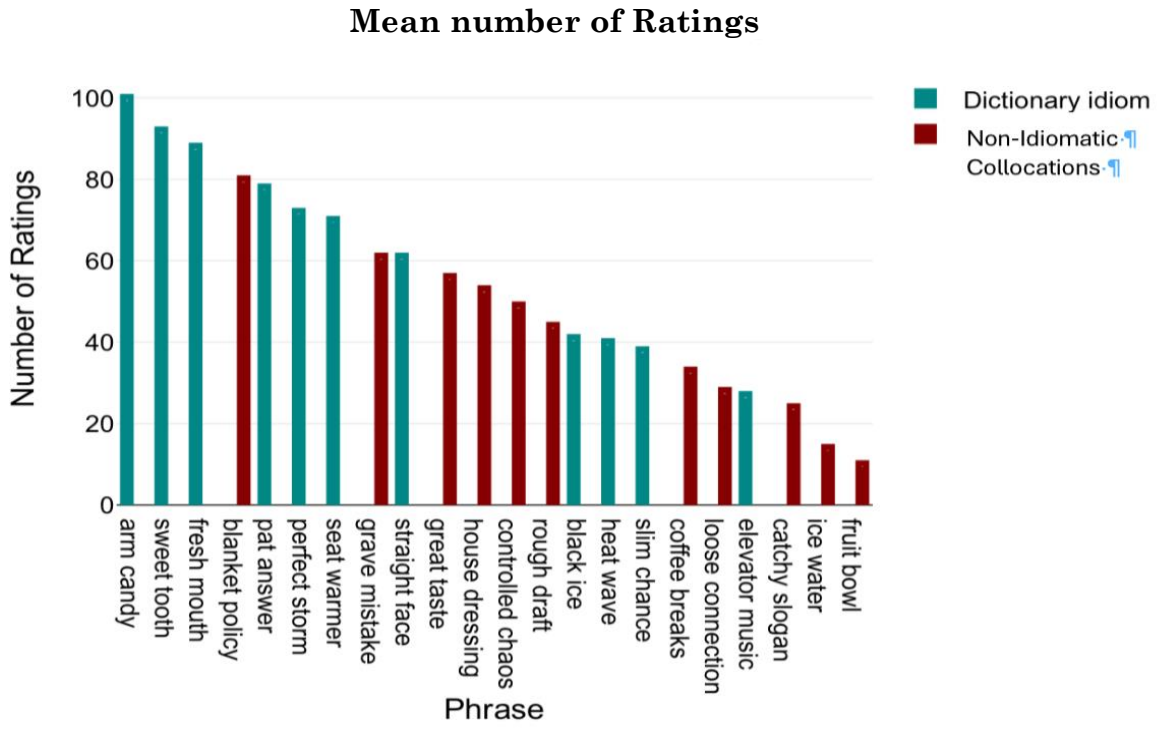


Figure 4. Gradated nature of observed responses. Figure 4 shows the gradation observed in the actual responses.

3.2.2. Mixed-effects logistic regression analysis

To address RQ1a, I conducted a reliability analysis using Fleiss' kappa, which revealed a poor to fair, unreliable, level of agreement among raters. This analysis provided valuable insight into the overall consistency of the ratings. However, on its own, Fleiss' kappa does not account for the observed variability. While it provides an overall measure of agreement it does not explain the sources of variability in the ratings.

To address these limitations, I conducted a mixed effects logistic regression analysis to evaluate the likelihood that dictionary idioms were more likely than non-idiomatic collocations to be rated as idiomatic. This analysis also considered the role of formulaicity, with participants and phrases as random effects.

This analysis adds to information provided by the kappa statistic by providing insights into how different factors influence idiomaticity ratings. Second, the mixed-effects method allows one to consider random effects, increasing the likelihood that the observed effects are due to the predictors of interest. Third, it allows complex interactions between variables to be modeled, which is crucial for understanding the nuances of rating reliability. By combining reliability analysis with regression, a more comprehensive understanding of the factors affecting rating consistency is obtained. While reliability analysis provides a snapshot of agreement, regression analysis delves deeper into the underlying cause of variability.

3.2.2.1. Analysis

To determine whether native speakers share general intuitions about perceived idiomaticity such that they are more likely to judge a phrase as idiomatic if it falls within the category of phrases taken from idiom dictionaries, a mixed-effects logistic regression was performed using the lme4 package in R (Bates, Maechler, Bolker & Walker 2015) in R (R Core Team 2014). Ratings of idiomaticity served as the dependent variable because participants were asked to evaluate perceived idiomaticity, irrespective of how they arrived at this judgement (i.e., irrespective of whether idiomaticity, a single property such as formulaicity, or multiple properties influence ratings). The small number of "unsure" ratings were removed from the data, leaving a binary, discrete, dependent variable and thereby justifying the use of logistic regression. The primary variable of interest was phrase type (idiom or equally formulaic non-idiomatic collocation), which was included as a predictor variable. Formulaicity (high PMI, medium PMI, and low PMI) was also included as a predictor variable to examine the potential influence of formulaicity on idiomaticity ratings. While PMI is, by nature, a continuous variable, scores were grouped into high, medium, and low bands for analysis (see section 3.1). To account

for baseline differences between raters, participants and phrases were included as random effects.

3.2.2.2. Results and discussion

The results suggest that phrase type is able to predict ratings, indicate no effect of formulaicity, and support the hypothesis that idiomaticity is a gradated construct. The analysis began with an evaluation of the descriptive statistics, which provided a foundational overview of the data by offering insights into the central tendencies, frequency, variability, and overall patterns.

The most relevant portion of the initial descriptive analysis was the response composition by condition. This considers the frequency with which participants selected “idiomatic” for idioms and equally formulaic non-idiomatic collocations. Table 3 summarizes the relevant descriptive statistics, including the proportion of responses rated as "idiomatic" for each condition and the probability a phrase would be rated as idiomatic or non-idiomatic.

Condition	Total ratings by condition	Total rated idiom	Total rated not Idiom	Portion rated idiom	Portion rated not idiom
Idiom	4962	3698	1264	0.639	0.332
Collocation	4939	1689	3250	0.248	0.727

Table 3: Descriptive statistics of response composition by condition.

The descriptive statistics are in line with the prediction that, while native speakers do not categorically distinguish between members of the class idiom, they appear to be aware of some sort of difference between idioms and non-idiomatic collocations. For dictionary idioms, the proportion of time that “idiomatic” was chosen was .64. This differed significantly from ratings of non-idiomatic collocations. For non-idiomatic collocations, the proportion of the time that “idiomatic” was chosen was

.25 (p<.001; see Figure 5). Thus, in this initial assessment, dictionary-defined idioms seem to influence speakers' judgments.

Following the descriptive assessment, a mixed-effects logistic regression was performed to further explore these patterns and to quantify the influence of predictor variables on the likelihood of a phrase being rated as idiomatic. A significant effect for phrase type was found. With dictionary idioms designated as the reference category for phrase type, equally formulaic non-idiomatic collocations were less likely to be rated as idiomatic than dictionary idioms (β=-2.18, z=-10.28, p<.001).

No effect of formulaicity on ratings was observed. High PMI was set as the reference category. There was no difference in the likelihood that a phrase would be rated as idiomatic if it had low PMI (β=-0.21, z=0.81, p=0.42) or medium PMI β=-0.02, z=0.01, p=0.92).

Random effects		
Group	Variance	Std.Dev.
Phrase	1.016	1.008
Participant	0.484	0.696

Fixed effects					
	Variable	Coefficient	Std. Error	z-value	p-value
	Intercept	1.41	0.214	6.584	4.58E-11
Phrase type	Collocation	-2.187	0.213	-10.284	2E-16
PMI	Low	-0.207	0.257	-0.805	0.421
	Medium	0.024	0.257	0.095	0.924

Correlation of fixed effects				
	Variable	Intercept	Collocation	PMI-Low
Phrase type	Collocation	-0.5		
PMI	Low	-0.542	0	
	Medium	-0.543	0	0.452

Table 4. Logistic regression analysis results.

To assess the significance of phrase type in predicting ratings, a likelihood ratio test was performed (Winters 2013). In this test, the full model was compared to a reduced model. In the full model, all variables were included (i.e., phrase type and PMI as fixed effects, with participant and phrase as random effects). In the reduced model, all variables were represented as in the full model except for phrase type, which was removed (i.e., PMI as a fixed effect with participant and phrase as random effects).

The likelihood ratio test indicated that the inclusion of phrase type significantly improved the model fit ($\chi^2(1) = 79.92$, $p < .001$). Additionally, model fit indices for the full models showed improved values compared to the reduced model, with lower AIC[20] and BIC[21] values and a higher log likelihood value, indicating a better fit (see Table 4 for values). These results suggest that phrase type is a significant predictor of ratings, and its inclusion in the model provides a significantly better fit to the data than the model without it. To confirm that formulaicity did not significantly impact ratings, a second likelihood ratio test was conducted. In this test, a reduced model which included phrase type as a fixed effect with participant and phrase as random effects was compared to the full model. While AIC, BIC, and log likelihood improved slightly, the likelihood ratio test indicated that the inclusion of PMI did not significantly improve the model fit, ($\chi^2(2) = .906$, $p < 0.636$), suggesting that PMI is not a significant predictor of ratings in the presence of phrase type (see Table 5 for values). Together, these tests indicate that phrase type significantly improves the model while PMI does not.

[20] The AIC, or Akaike information criterion, is a measure used to evaluate the goodness of fit of a statistical model while penalizing for model complexity. Lower AIC values indicate a model that balances fit and simplicity more effectively, making it useful for comparing models where the model with the lowest AIC is preferred (Cavanaugh & Neath 2019).

[21] The BIC, or Bayesian Information Criterion, is similar to AIC but has a stronger penalty for model complexity, particularly in larger datasets. Like AIC, lower BIC values indicate a better-fitting model. However, compared to AIC, BIC tends to favor simpler models (Cavanaugh & Neath 2019).

Likelihood Ratio Test for Phrase Type							
	AIC	BIC	LogLik	Deviance	Chisq	df	p-value
Reduced Model	14528	14628	-7285.1	14570			
Full Model	14504	14558	-7245.1	14490	79.92	1	2.2e-16***
Likelihood Ratio Test for Formulaicity (PMI)							
	AIC	BIC	LogLik	Deviance	Chisq	df	p-value
Reduced Model	14501	14539	-7245.6	14491			
Full Model	14504	14558	-7245.1	14490	0.9063	2	0.6356

Table 5. Model fit indices for the reduced and full models.

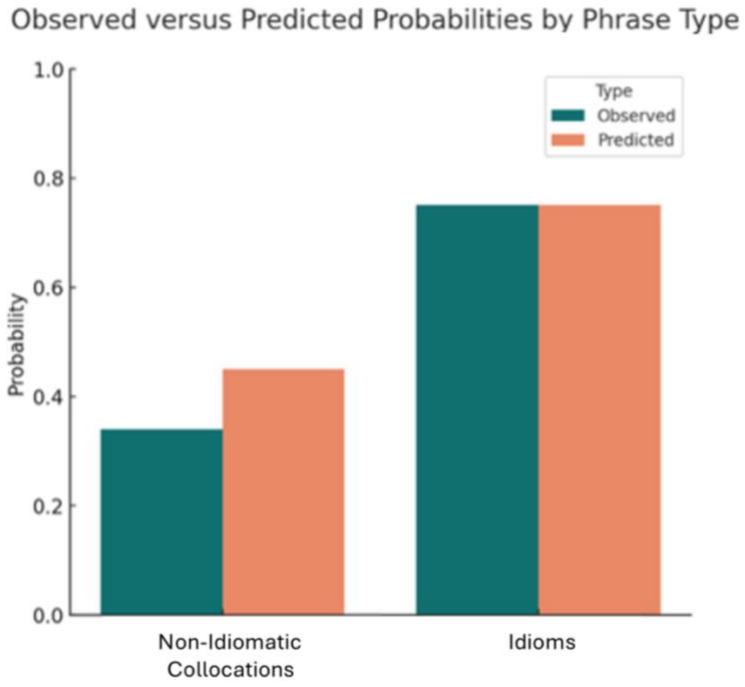


Figure 5. Observed and predicted ratings probabilities by condition (idiomatic versus non-idiomatic).

Overall, the results of the regression indicate that phrase type is a significant predictor of ratings, regardless of how formulaic a phrase is. The descriptive measures indicate a tendency to rate idioms as idiomatic more often than non-idiomatic collocations, as evidenced via a higher frequency of "idiomatic" ratings for idioms as compared to equally formulaic non-idiomatic collocations. The regression analysis corroborated this, finding that a phrase was significantly more likely to be rated as idiomatic if it was an idiom than if it was a collocation. However, this relationship was not absolute. If idioms exist on a continuum, it may be that phrases with particular combinations of features make an idiom seem more idiomatic, resulting in reliable differentiation for specific types of idioms. Similarly, non-idiomatic collocations that least express these properties may be more reliably rated as non-idiomatic. Experiment 2 explored relationships between features such as the degree of formulaicity, distributed semantics, partial literality, and dual meaning to determine if certain combinations of properties can predict which phrases are construed as idiomatic.

3.4. Conclusions, limitations, and future directions

Experiment 1 addressed RQ1a and 1b, seeking to determine whether native speakers share intuitions about idiomaticity by investigating the impact of idiomaticity on how we think about multiword phrases. The results of experiment 1 reveal a nuanced category, reflecting the complex nature of the cognitive construct idiomaticity. The low degree of agreement at the item level supports the hypothesis that there is no cleanly delineated class idiom whose members are either "idiomatic" or "not idiomatic" and about which native speakers share intuitions. If idioms are stipulated in the mental lexicon as "idiomatic", native speakers are not consciously aware of the special status afforded to these phrases. However, this does not mean that there is no mental category of idiomaticity. Consideration of the frequency of

ratings by type as well as the regression analysis indicated that native speakers share an awareness of some sort of difference between idioms and non-idiomatic collocations.

These results address long-standing assumptions within psycholinguistics, indicating that a single criterion definition of idiomaticity, in which idiomaticity is synonymous with nondecompositionality, is insufficient. Instead, the findings are in line with linguistic work demonstrating that the line between idioms and other kinds of lexicalized phrases is not always clear (cf. Kamp & Partee 1995, Nenonen 2007, Wary 2013), indicating that further investigation into the role of properties of idioms is needed as a prototype-based approach may be a more appropriate way to account for idiomaticity than a binary classification (cf. Penttila 2010). While experiment 1 showed that, as a class, idioms are not reliably differentiated from equally formulaic, non-idiomatic collocations, it is unable to account for the general pattern of phrase type judgments, which showed significantly more ratings of "idiomatic" for phrases in the dictionary idiom condition and significantly more ratings of "non-idiomatic" for phrases in the equally formulaic non-idiomatic collocation condition. Additionally, it is unable to account for the observed response gradation. If idioms are a continuous construct, it may be that certain properties, such as distributed idiomatic meaning, partial literality, and dual meaning may be able to account for this variation. It is this eventuality that experiment 2 seeks to address. Specifically, experiment 2 evaluates the relationships between hypothesized idiom properties and ratings of perceived idiomaticity, examining whether these properties predict which phrases will be reliably recognized as idiomatic. Such an approach would not only account for the observed variability in ratings but would also provide a framework for future research into how idiomaticity is represented and processed in the mental lexicon. Thus, experiment 2

builds on the findings of experiment 1 to further elucidate the cognitive construct of idiomaticity and the properties that define its prototypical features.

In addition to indicating that idiomaticity is not a discrete category, the findings indicated that experimental study of idioms should be approached with extreme care. A novel feature of Fleiss' kappa is that it does not assume that experimenter-defined conditions (eg., "idiom" and "non-idiomatic collocation") are "correct". By this, I mean that the analysis does not compare a dependent variable against an experimenter-defined independent variable to determine the degree to which the dependent variable differs from the experimenter-defined variable. Instead, it considers only the degree to which participants differ from one and another on a given judgement. Metrics such as this are vital. In light of the findings that native speakers do not reliably recognize idioms, one must otherwise wonder what it is that current experimental work is testing. If phrases are grouped into an "idiom" condition or "non-idiom" condition based on a researcher's analysis but these categorizations are not shared by participants, we must ask whether the task is actually testing idiomaticity.

One methodological strength of this experiment was its ability to directly test conceptualization rather than relying on a proxy measure such as time in a reaction time experiment. However, future work should consider the relationship between the present findings and those from online tasks due to inherent limitations of using a metalinguistic judgment task to study idiomaticity. Metalinguistic tasks are valuable for eliciting explicit judgments about linguistic constructs and it is generally accepted that native speaker analytic abilities and intuitions are correlated with online processing (cf. Tabossi et al. 2008). However, this has not been validated for phrase type recognition and should not be assumed as metalinguistic judgements do not necessarily reflect the implicit, online processing

mechanisms involved in natural language comprehension and production. Metalinguistic tasks require participants to consciously analyze and evaluate linguistic forms, which may engage cognitive processes distinct from those used during spontaneous language use. For example, participants may rely on post-hoc reasoning or draw on their familiarity with conventional labels (e.g., "idiom" or "non-idiom") rather than intuitive processing. This distinction is particularly relevant in light of the findings that native speakers often exhibit low agreement in idiomaticity ratings, as such variability could partly reflect the cognitive demands and introspective nature of the task itself. Future research should complement metalinguistic tasks with online methodologies, such as reaction time, eye-tracking, or EEG, to better understand how idiomaticity is processed in real time.

Overall, the results of experiment 1 address long-standing assumptions regarding the nature of the class idiom. However, these results raise a number of questions. In particular, they are unable to provide insight into factors that motivate native speaker intuitions of perceived idiomaticity, which are necessary to account for the observed variance in ratings. When the observed continuous nature of idiomaticity is considered in the context of prior work demonstrating the impact of properties on idiom conceptualization and comprehension, the findings indicate that further investigation into the role of idiom properties is needed. It is this point that experiment 2 addresses.

CHAPTER IV

EXPERIMENT 2

Experiment 1 considered the role of idiomaticity in how we think about idioms by investigating whether native speakers are able to reliably recognize idioms, challenging the idea that there is a discrete class idiom at the conceptual level. The results indicate that this is not the case. Instead, they suggest that the line between an idiom and a non-idiom is blurry and that the class cannot be defined by the single criterion of noncompositionality. Additionally, the findings suggest that idiomaticity is continuous, with speakers sharing intuitions for certain phrases at a level that is highly reliable while agreement for others is lower than chance. Given the continuous nature of idioms, the class idiom may be better accounted for by a prototype-based approach, in which multiple properties, such as formulaicity, distributed idiomatic meaning, partial literality, and dual meaning are associated with idiomaticity. In such an approach, a phrase displaying more properties may be associated with a higher degree of perceived idiomaticity while a phrase displaying fewer properties would be associated with a lower degree of perceived idiomaticity. It is this eventuality that experiment 2 considered.

The goal of experiment 2 was to evaluate the relationships between the four hypothesized idiom properties and perceived idiomaticity to determine whether these properties are associated with prototypical idiomaticity. Specifically, experiment 2 builds on the findings of experiment 1, seeking to better understand the cognitive determinants of idiomaticity by addressing the second research question (**RQ2**), which asks: to what degree are formulaicity, distributed idiomatic

meaning, partial literality, and dual meaning able to predict which phrases will be reliably recognized as idiomatic? To this end, experiment 2 presents a novel analysis of idiom conceptualization that evaluates the relationship between four hypothesized idiom properties and ratings of perceived idiomaticity, seeking to determine which, if any properties describe the cognitive construct of idiomaticity and which serve as a prototypical feature of the class.

While some have challenged the psychological validity of idiom properties (cf. Bobrow & Bell 1973, Swinney & Cutler 1979, Cutler 1982, Cacciari & Tabossi 1993, Glucksburg 2001, Vulchanova et al. 2015, Qualls & Harris 2018, Milburn 2020), prior work has shown that a number of properties are important to idiom conceptualization and comprehension. As previously mentioned, comprehension refers to how one engages with an idiom in real-time, drawing on pre- or subconscious processes that allow for understanding without necessarily requiring full analysis of individual components. Conceptualization, in contrast, involves metalinguistic reflection and is concerned with forming, organizing, and structuring knowledge about concepts and mental categories, as well as analyzing idioms by their individual properties and their broader categorization within figurative language. The most widely-studied idiom properties are distributed idiomatic meaning (whether or not a portion of the meaning of a phrase is associated with each word), partial literality (whether or not one word contributes a meaning within the phrase that it can independently convey outside of the phrase), dual meaning (whether or not a phrase has a plausible literal and idiomatic interpretation), and formulaicity (how likely it is that one will encounter constituent words in an idiomatic construction rather than on their own or in another construction). While the impact of these properties on idiom comprehension has received more attention, idiom properties have also been shown to impact idiom conceptualization, with prior

work collecting ratings of properties to determine the degree to which native speakers share intuitions regarding properties, the impact of certain properties on the ratings of other properties, and the relationships between idiom conceptualization and comprehension (cf. Gibbs & Gonzolas 1985, Gibbs & Nayak 1989, see also Titone & Connine 1994a, Tabossi et al. 2008, Nordmann et al. 2013). However, the relationship between properties and the cognitive construct of idiomaticity remains assumed but untested, as no study has directly examined whether ratings are directly reflective of the conceptualization of a phrase as a member of the class idiom. Thus, these properties may be the ideal place to begin to decompose cognitive idiomaticity. This is not to say that the properties of distributed idiomatic meaning, partial literality, dual meaning, and formulaicity are in any way sufficient to fully describe the class idiom. Given the amount of variation between phrases within the class, it is likely that a number of additional properties are also associated with idiomaticity, particularly (in-)flexibility and transparency. The chosen properties provide a starting point as they are the most widely studied and, at present, are thought to be the most impactful. Until we gain a more precise understanding of the mental construct of idiomaticity, we will remain at an impasse in advancing this line of research.

Additionally, while relationships between certain properties, such as distributed idiomatic meaning and partial literality have been established, work considering relationships between all three subjective properties is rare (c.f. Titone & Connine 1994a). To date, no work has modeled relationships between the three subjective properties, the objectively measured formulaicity, and perceived idiomaticity. This knowledge is critical to advance our understanding of idiom conceptualization and comprehension as it is impossible to model a construct we cannot define.

I argue that relationships exist between properties and idiomaticity such that the presence of certain properties may make one phrase (e.g., *elbow grease*) seem more idiomatic than another (e.g., *best shot*). If this is the case, the inherently multidimensional nature of idiomaticity must be factored into any complete model of cognitive processing.

To investigate the impact of properties, experiment 2 employed methods that relied on both objective measures and subjective judgments provided by participants. First, to investigate formulaicity, it utilized the pointwise mutual information scores calculated prior to experiment 1. Next, it used three categorization tasks to collect native speaker judgments of distributed idiomatic meaning, partial literality, and dual meaning. Each task included the same 144 phrases used in experiment 1. Participants read each phrase then decided whether the final phrase displayed a property (e.g., *slim chance* and *best shot* have distributed idiomatic meaning) or not (e.g., *arm candy* and *elbow grease* do not have distributed idiomatic meaning, see Appendix III).

A mixed effects logistic regression analysis was used to evaluate the relationships between properties and perceived idiomaticity to determine whether properties influence how we think about idioms. Findings in which no relationships between phrase type recognition and the investigated properties exist would argue against a prototype-based approach, at least one in which these properties are treated as reflective of idiomaticity. However, I predict that properties will individually and collectively impact the recognition of a phrase as an idiom. Such a finding would indicate that properties, even some lexical-level properties, play a role in how we think about idioms and must be considered in experimental work. These properties are expected to impact idiomaticity judgments because they moderate the degree to which a phrase differs from a non-idiom. Two key examples of this are in the degree

of analyzability and the degree of distinctness, or how much a phrase differs semantically and syntactically from what would be expected of a well-formed literal version of the phrase. Thus, idiom prototypicality is a function of the degree of difference between an idiom and a non-idiom.

This can be seen in the correlation between distributed idiomatic meaning and flexibility where information about allowable instantiations can be determined by analyzing the individual parts. By contrast, phrases with no distributed idiomatic meaning are more fixed in nature. Generally speaking, syntactic information about tense, aspect, and modality can be derived from phrases with distributed idiomatic meaning while such information must be memorized for phrases with no distributed idiomatic meaning as it is not unusual for there to be a mismatch (e.g., *pushing up daisies* (activity), which means "to be dead" (state)). This differs from the impact of a property like partial literality, which also affects analyzability, but primarily at the lexical level. When one word retains its non-idiomatic meaning, it becomes at least partially analyzable. This is not the same as decompositionality – noncompositional phrases can be partially literal. Instead, this type of analyzability reduces the distance between the meaning of the individual constituents and their non-idiomatic meaning. Importantly, partially literal phrases are not necessarily transparent. It is simply that at least one word conveys a familiar (non-idiomatic) meaning. During comprehension, this may be particularly advantageous as a word that retains its literal meaning within an idiomatic phrase may be activated more strongly than other words, aiding in retrieval while conflicting meaning between all words in a phrase that is not partially literal delays the point at which one can commit to an idiomatic meaning.

I predict that a phrase is most likely to be rated as idiomatic if it has no distributed idiomatic meaning, is not partially literal (i.e., neither word conveys a non-idiomatic

meaning), and has no dual literal meaning, such as *arm candy, sweet tooth,* and *sack time*). Conversely, I predict that a phrase is least likely to be rated as idiomatic if it has distributed idiomatic meaning, is partially literal, and has dual meaning (e.g., *moving target, big picture, track record*). Judgements of idiomaticity for different combinations of properties will vary between these. These predictions are based on idiom prototypicality, where the more prototypical an idiom is, the more it stands out as different from a non-idiom. For example, a phrase with distributed idiomatic meaning is analyzable, even though the meaning associated with constituent words is not literal. From a cognitive standpoint, this makes the phrase more similar to a non-idiom than a phrase with no distributed idiomatic meaning because, each word in a phrase with distributed idiomatic meaning contributes a portion of the phrase's overall meaning. Similarly, phrases that are partially literal may seem less prototypically idiomatic than phrases that are not partially literal. This is because, in a partially literal phrase, one word is used in the same way inside the phrase as it is outside of it, which makes the phrase at least partially analyzable, may add a degree of transparency, and creates an illusion of more free combinatorics.

Like distributed idiomatic meaning and partial literality, the lack of a plausible dual literal and idiomatic meaning is associated with a higher degree of prototypical idiomaticity. However, the reason for this association differs. Dual meaning is a phrasal-level property, unlike the lexical-level partial literality or the primarily lexical distributed idiomatic meaning. The presence of distributed idiomatic meaning and partial literality makes a phrase more analyzable because meaning can be associated with individual words. Instead of constituent analyzability, dual meaning relates to prototypicality via the comparison of meanings between a non-idiomatic and idiomatic interpretation. During comprehension, dual meaning can

initially slow processing. This is because when a phrase has plausible dual literal and idiomatic meaning, both representations are activated. One must determine which interpretation is appropriate in a given context then suppress the meaning of the competing interpretation. However, dual meaning has been shown to facilitate comprehension when considered on a larger time scale (Libben & Titone 2014, Morid 2021). This is particularly true for phrases with a clear relationship between their idiomatic and non-idiomatic interpretations (Libben & Titone 2014). In behavioral experiments, such as ratings tasks, phrases with no dual meaning are often more reliably rated (cf. Gibbs & Nayak 1989, Nordmann et al. 2013) as relatedness between interpretations can interfere with judgements (e.g., the idiomatic interpretation of *big picture* is a clear metaphoric extension of the non-idiomatic interpretation). This interference is predicted to extend to perceived idiomaticity, reducing the likelihood that phrases with dual meaning will be rated as idiomatic as compared to those with no dual meaning.

4.1. Methodology

In this section, I describe the methodology for the distributed idiomatic meaning, partial literality, and dual meaning tasks. It begins with a brief reminder of the phrase type recognition task (experiment 1), Next, it discusses task-specific differences, namely the judgments solicited in each task. It concludes with a description of the participants and of the collected data.

4.1.1. Materials

Three ratings tasks were used to collect native speaker intuitions regarding idiom properties. The data was then analyzed using a mixed effects logistic regression to identify potential relationships between properties and perceived idiomaticity. Results from the ratings tasks were compared with the results from experiments 1

and PMI scores to determine whether these properties were able to explain what makes a phrase more recognizably idiomatic.

The ratings tasks used in experiment two were similar to those used in experiment 1. All tasks were based on the same underlying task design as that employed in experiment 1, collecting judgements for the same phrases embedded in the same sentences and differing only in the requested judgement. Three Qualtrics tasks were created; one collecting ratings of distributed idiomatic meaning, one collecting ratings of partial literality, and one collecting ratings of dual meaning. These tasks, including how phrases were categorized, are explained below. Each task began with the same consent form and screening questions. Participants who consented to participate and pass the screening questions read task instructions, completed practice ratings and received feedback. After completing the practice ratings, participants were taken to the task, where they rated the same 144 sentences used in experiments 1-3. Like experiment 1, two attention check questions were included in the distributed idiomatic meaning, partial literality, and dual meaning tasks. Each task took about 45-60 minutes to complete.

Careful consideration was given to task instructions. Adoption of previously used experimental instructions was strongly considered in order to be as consistent as possible with prior work. However, it was decided that strict adherence to this option could introduce potential confounds. Instead, multiple iterations of the task instructions were tested and refined, optimizing clarity and task simplicity while ensuring instrument validity.

Three pilot studies were conducted. The purpose of Pilot study 1 was to assess instruction clarity, the cognitive demands of the task, and completion time. Pilot 2 considered the user experience, seeking to optimize task length, interface display,

and response option efficiency to improve response quality and completion rates by minimizing extraneous cognitive demands. In pilot 3, final refinements were tested to ensure task effectiveness and fidelity. In pilot study 1, participants (n=6) rated a subset of test phrases for all three subjective properties (distributed idiomatic meaning, partial literality, and dual meaning). Task instructions were consistent with Gibbs et al. (1989), Titone and Connine (1994a), and Libben & Titone (2008) as were response options. For example, in the distributed idiomatic meaning task, participants rated a phrase as "decompositional" or "noncompositional" (Gibbs & Nayak 1989, Titone & Connine 1994a, Libben & Titone 2008, see also Nordmann et al. 2014). In the partial literality task, participants determined whether a phrase was "partially literal" or "not partially literal" (Titone & Connine 1994a, Libben & Titone 2008). In the plausible dual literal and idiomatic meaning task, participants determined whether phrases could be used "idiomatically only", "literally and idiomatically", or "literally only" (Titone & Connine 1994a, Libben & Titone 2008, Tabossi et al. 2011, Nordmann et al. 2014). Overall, the results of pilot study 1 provided initial support for the predictions. However, it identified important points that needed to be addressed. These included the need for additional examples, a re-wording of unclear or ambiguous instructions, and re-naming and simplification of response options. Additionally, it was clear that the task was too cognitively taxing, which led to a marked decrease in response quality before the halfway point of task completion. To address this problem, the task, which collected ratings of all properties, was split into three separate tasks such that only one property was rated in each task. Pilots 2 and 3 further refined these changes. Specifically, the main goal of pilot study 2 was to gather feedback for further refinement of task instructions so as to remove any potentially confusing or misleading phrasing, to optimize the response format by reducing the visual burden of cluttered pages, and to verify a reduction in task design-induced cognitive demands resulting from

splitting the single task into multiple separate tasks, each gathering ratings for only one property. In pilot study 2, participants provided responses to a subset of the phrases included in pilot study 1 then took part in a focus group. In the focus group, participants were asked to describe their experience, to discuss what they liked and did not like about the task (e.g., instructions, practice phrases and response explanations, layout, etc), and to provide live feedback during an interactive design session in which suggested changes related to the user experience were implemented and the alternatives discussed. Thus, pilot study 2 provided valuable feedback from a user experience perspective. However, as intended, data collected from pilot study 2 was not robust enough to assess even exploratory results due to the small sample size and small subset of included phrases. As a result, a final pilot study was conducted. The purpose of pilot study 3 was twofold. First, it tested the effectiveness and fidelity of final refinements. Second, a sufficient amount of data was collected so as to test the hypotheses. The results of pilot study 3 were included in four accepted grant proposals, which funded data collection. Two of these grants were awarded by the Institute of Cognitive Science (ICS), one was awarded by the Center to Advance Research and Teaching in the Social Sciences (CARTSS), and one was awarded by the University of Colorado Boulder Department of Linguistics. Additionally, the University of Colorado Boulder Department of Linguistics provided funding for the initial data collection performed in all pilot studies.

The following subsections discuss task-specific points, including the instructions and task experience. Crucially, no tasks included the word “idiom”. The goal of this experiment was to test the theory that distributed idiomatic meaning, partial literality, and dual meaning are reflective of the mental construct of “idiomaticity”, not to test participants’ ability to recognize an idiom, which would conflate these

tasks with that of experiment 1 and invalidate the findings. Additionally, the term "literal" was not used in the distributed idiomatic meaning task or the partial literality task. While "literal" appeared in the dual meaning task, its use was vital. It is fully recognized that "literal" is not the opposite of "idiom". When possible, this term was not used. However, it was necessary to avoid use of the term "idiomatic".

4.1.1.1. Distributed idiomatic meaning task

In the distributed idiomatic meaning task, participants determined whether phrases had "additive meaning" or "non-additive meaning" (see Appendix III for an example of the task). A phrase has "additive meaning" if each word contributes an identifiable portion of the meaning of the phrase. A phrase has "non-additive meaning" if each word does not contribute an identifiable portion of the meaning of the phrase. For example, participants were told that *trash talk* and *blue car* both have additive meaning. This is because in the phrase *trash talk*, *trash* is associated with "insults" or "boasting" and *talk* is associated with speech. Similarly, in the phrase *blue car*, *blue* is associated with something that is blue and *car* is associated with something that is a car. Conversely, the phrases *lead foot* and *loan shark* have non-additive meaning. This is because in the phrase *lead foot*, neither word is associated with a meaning like "driving" or "very fast". In the phrase *loan shark*, *loan* is associated with a kind of *loan*. However, no identifiable meaning is associated with *shark*.

The response options "additive" and "non-additive" were chosen to avoid the possibility of participants simplifying the task to one in which their responses were based, not on an assessment of distributed meaning, but on whether a phrase was idiomatic or non-idiomatic. By not introducing terms such as "idiomaticity" or "literality", participants were able to complete the task as instructed.

4.1.1.2. Partial literality task

In the partial literality task, participants considered the meaning of individual words in a phrase and determined if no words, one word, or two words had a meaning in the phrase that they can also have outside of the phrase. For example, the phrase *cold turkey* is not really about something that is cold or that is a turkey, so neither word has the same meaning inside the phrase as it can outside of the phrase. In the phrase *snail mail*, *mail* does refer to a kind of mail but it does not involve snails. So, one word in *snail mail* has the same meaning inside of the phrase as it can outside of it. In the phrase *blue car*, *blue* refers to a color and *car* refers to a kind of motorized vehicle. So, both words have the same meaning in this phrase as they can outside of it.

To rate partial literality, participants used the response options of "both typical", "one typical" and "no typical" where "typical" refers to the conveyance of the same meaning within a phrase as outside of it. For example, "both typical" was explained as follows:

When both words in a phrase have the same meaning within the phrase that they generally have when used outside of it, we would say that **both** words are used in a **typical way**.

Similar to the distributed idiomatic meaning task, the response options and task instructions avoided words such as "literal" or "idiomatic" that could potentially allow for a participant to reduce the task to one in which ratings were conflated with judgments of idiomaticity.

4.1.1.3. Plausible dual literal and idiomatic meaning task

In the dual meaning task, participants determined whether a phrase was used literally or non-literally in the provided context then decided if that phrase could

also be used in the other way. For example, the phrase *blue car* can be used literally only and the phrase *sweet tooth* can be used non-literally only. So, these phrases have only one meaning. Conversely, the phrases *thin ice* and *bad apple* have plausible literal and non-literal meanings since they can be used both ways. Thus, participants selected from three ratings options: literally only, non-literally only, or both literally and non-literally.

While words such as "literal" and "non-literal" were used in this task, the response options avoided potential overlap with the phrase type recognition task by not introducing the concept of idiomaticity, ensuring that this task tested the hypothesized relationship between dual meaning and idiomaticity without collecting ratings of idiomaticity.

4.1.2. Participants

Participants were recruited from Prolific and CU Boulder. Each Prolific participant received $12 as compensation for their time. CU Boulder students were recruited using flyers, word-of-mouth, and the linguistics department extra credit pool in the fall of 2023 and spring of 2024. Students participating for extra credit were enrolled in LING 1000: Language and U.S. Society, LING 2000: Introduction to Linguistics, or LING 3100: Language Sound Structures. Extra credit students were given the option of receiving extra credit in their linguistics course or receiving a $12 e-gift card to Amazon or Starbucks. All other CU participants who were not enrolled in an extra-credit granting class received a $12 e-gift card to Amazon or Starbucks.

A power analysis indicated that 60 completed tasks of each type (distributed idiomatic meaning, partial literality, and dual meaning) were needed to perform this analysis. To help combat the chronic problem of underpowered studies, data

from 110-120[22] participants were collected for each task, with the goal of including data from at least 100 participants for each task type in the final analysis.

The dual meaning ratings task had the smallest number of participants, establishing the inclusionary totals for all other variables. In total, 111 participants completed the dual meaning task, 71 of which were members of the Prolific community and 40 of which were CU students. Four Prolific participants and four CU students were excluded from the final analysis due to failed attention checks, resulting in data from 103 participants for analysis.

112 participants completed the partial literality task. Of these, 62 participants were members of the Prolific community and 50 were CU students. Two Prolific participants and three CU students were excluded from the final analysis due to failed attention checks, resulting in data from 107 participants for analysis.

119 participants completed the distributed idiomatic meaning task, 31 of which were members of the Prolific community and 88 of which were CU students. Seven CU students were excluded from the final analysis due to failed attention checks. All Prolific participants were included because all successfully completed the attention checks. This resulted in data from 112 participants for analysis.

4.2. Results and discussion

To evaluate the relationship between properties and perceived idiomaticity (RQ2), ratings of distributed idiomatic meaning, partial literality, and dual meaning for dictionary idioms collected in experiment 2 as well as objectively calculated formulaicity and the idiomaticity ratings of dictionary idioms collected in

[22] This overage was established on the assumption that 1-10 participants per study would need to be excluded from the final analysis for reasons such as failure to complete the entire task or failed attention checks.

experiment 1, were analyzed using a mixed-effects logistic regression. The results indicate that three out of the four idiom properties impact perceived idiomaticity.

4.2.1. Analysis

The mixed effects logistic regression was completed using generalized linear modeling with a logit link function in IBM SPSS Statistics (Version 28) and the lme4 package (Bates, Maechler, Bolker & Walker 2015) in R (R Core Team 2014) to determine whether distributed idiomatic meaning, partial literality, dual meaning, and formulaicity are able to predict ratings of idiomaticity, and to test whether certain combinations of properties are associated with a higher degree of idiom prototypicality. In this analysis, idiomaticity ratings collected in experiment 1 served as the dependent variable. In the experiment 1 phrase type recognition task, participants rated the same 144 phrases that were used in the distributed idiomatic meaning, partial literality, and dual meaning tasks as "idiomatic", "non-idiomatic", or "unsure". Participants were asked to use "unsure" no more than 5 times and these responses, which constituted less than 2% of the total responses for this task, were removed from the analysis.

Ratings of distributed idiomatic meaning, partial literality, dual meaning, and formulaicity served as fixed effects. The main effect of each was considered to determine the degree to which each property independently predicts idiomaticity, while controlling for the other predictors in the model. Additionally, participants and phrases were included as random effects to account for systematic variance related to these groups but unrelated to the investigated variables. A nested structure was used to reflect the independent nature of the fixed effects such that different participants rated distributed idiomatic meaning, partial literality, and dual meaning. To assess the individual contribution of properties to the model, Wald chi-square was used. Wald chi-square directly assesses the significance of

each predictor variable in a model and is preferable to log likelihood when assessing the relationship between a specific independent variable and the dependent variable (Menard 2002, Agresti 2012).

Overall, the analysis allowed for the evaluation of relationships between each property and idiomaticity, determining whether each plays a significant role in perceived idiomaticity such that it can be used to predict whether a phrase will be rated as idiomatic or not. While relationships between certain independent variables as well as the relationship between an independent variable and perceived idiomaticity have been studied previously, this work considers the relationship between multiple properties and perceived idiomaticity within a single analysis, increasing our understanding of the mental construct of idiomaticity. Specifically, it provides the first experimental account of the definition of an idiom by testing the ability of properties to describe this mental category. Further, it allows for the creation of categories reflective of native speaker intuitions, which will be used to create a continuum of idiomaticity. By considering tacit knowledge shared across participants as well as at multiple levels of PMI (low, medium, and high)[23], this work removes reliance on assumed shared category boundaries between the experimenter and participants. These results not only offer a more accurate reflection of the cognitive definition of an idiom, demonstrating the relationship between properties and idiomaticity in how we think about idioms, but also challenge traditional categorical approaches, providing a framework for understanding idiomaticity as a complex, gradient phenomenon.

To ensure an equal sample size across tasks, data from only 103 participants for each of the four ratings tasks (idiomaticity, distributed idiomatic meaning, partial

[23] PMI values covered a full spectrum of continuous values. Continuous values were forced into the discrete categories of low, medium, and high for this analysis.

literality, and dual meaning) were included in this analysis as this number was equal to the smallest task sample size. After ensuring that no tasks showed extreme deviance[24] in the recorded ratings, all collected dual meaning tasks (n=103) were retained for inclusion in the final analysis. Four participants from the partial literality sample (n=107) were randomly dropped to reduce the number of analyzed tasks to 103. Nine participants were dropped from the distributed idiomatic meaning sample (n=112) to reduce the number of analyzed tasks to 103. Three were selected for exclusion because of responses that deviated from the mean by 1 standard deviation and because they left self-reflective comments in the feedback section, saying things such as, "This task was much more difficult than I anticipated, but I am just too tired. Very engaging though! Thank you." The remaining participants were selected at random for exclusion. Of the 106 idiomaticity ratings tasks included in the experiment 1 analyses, three were randomly dropped from the final analysis.

4.2.1.1. Mixed effects logistic regression

Overall, the results indicate that distributed idiomatic meaning, partial literality, and dual meaning are able to predict ratings of idiomaticity, suggesting that these properties describe the class idiom and may serve as indicators of prototypical idiomaticity. One property, formulaicity, did not predict ratings of idiomaticity, suggesting that it may serve as a non-differentiating characteristic of idiomaticity.

[24] Extreme deviance was defined as responses more than 2 standard deviations from the mean. It was calculated using z scores for the purposes of determining the degree to which each participant's responses deviated from the overall task mean. First, the task mean was calculated by averaging the responses across all participants for each task. This mean served as the central tendency for the task. Next, the population standard deviation (stdev.p), which reflects the overall variability of responses across participants, and the sample standard deviation (stdev.s), which reflects the variability within each participant's responses, were calculated. The population and sample standard deviations were used to calculate the relative distance of each participant's responses from the task mean, capturing variability within individual ratings. Next, the z-scores were calculated using the task mean and the population standard deviation. This was accomplished by first determining the task means. Then, the following formula was used to calculate the z-scores: z = (participant response-task mean)/(task population standard deviation).

Additionally, no correlations were observed between properties. This can be seen in Tables 6-8, which summarize the relationships between properties and perceived idiomaticity. The following subsections discuss the relationship between each property and perceived idiomaticity, examining individual property contributions to the full model and evaluating the significance and impact of each property.

Covariances of Parameter Estimates for Fixed Effects					
Parameter	Parameter Explanation	Coefficient	Std. Error	z value	p-value
(Intercept)	(Intercept)	-1.92	0.099	19.49	<.001
+DIS	Has distributed idiomatic meaning.	.[a]	.	.	.
-DIS	Does not have distributed idiomatic meaning.	0.397	0.068	5.831	<.001
-PL_TwoLit	Not partially literal. Both words contribute a meaning they can independently convey outside of a test phrase.	.[a]	.	.	.
+PL_OneLit	Partially literal. One word contributes a meaning it can independently convey outside of a test phrase.	0.584	0.077	7.509	<.001
-PL_NoLit	Not partially literal. Neither word contributes a meaning it can independently convey outside of a test phrase.	0.633	0.101	6.257	<.001
-DM_LitOnly	No dual meaning. Has a plausible non-idiomatic interpretation only.	.[a]	.	.	.
+DM_Both	Has dual meaning.	0.244	0.076	3.226	0.001
-DM_IdOnly	No dual meaning. Has a plausible idiomatic interpretation only.	0.769	0.097	7.951	<.001
PMI-L	Low PMI.	.[a]	.	.	.
PMI-M	Medium PMI.	0.049	0.082	0.597	0.551
PMI-H	High PMI.	-0.057	0.088	0.322	0.401

a. Reference level. Set to zero because this parameter is redundant.

Table 6. Logistic regression parameter estimates.

Correlations of Parameter Estimates												
	(Intercept)	PMI-L	PMI-M	PMI-H	-PL_Two Lit	+PL_One Lit	-PL_Two Id	-DIS	+DIS	-DM_Lit Only	+DM_Both	-DM_Id Only
(Intercept)	1.000	-0.442	-0.395	.[a]	-0.367	-0.551	.[a]	-0.347	.[a]	-0.318	-0.443	.[a]
PMI-L	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]
PMI-M	-0.395	0.456	1.000	.[a]	-0.027	0.065	.[a]	-0.048	.[a]	0.004	-0.017	.[a]
PMI-H	0.543	1.000	0.564	.[a]	0.032	0.053	.[a]	-0.042	.[a]	0.043	0.098	.[a]
-PL_Two Lit	-0.367	0.022	-0.027	.[a]	1.000	0.482	.[a]	-0.029	.[a]	-0.044	-0.016	.[a]
+PL_OneLit	-0.551	0.050	0.065	.[a]	0.482	1.000	.[a]	0.011	.[a]	-0.021	0.018	.[a]
-PL_Two ID	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]
-DIS	-0.347	-0.040	-0.048	.[a]	-0.029	0.011	.[a]	1.000	.[a]	-0.015	-0.025	.[a]
+DIS	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]
-DM_Lit Only	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]	.[a]
+DM_Both	-0.443	0.096	-0.017	.[a]	-0.016	0.018	.[a]	-0.025	.[a]	0.421	1.000	.[a]
-DM_Id Only	-0.239	0.030	0.004	.[a]	-0.044	-0.021	.[a]	-0.015	.[a]	1.000	0.421	.[a]

[a] Reference level. Set to zero because this parameter is redundant.

Table 7. Correlation between variables indicating no multicollinearity.

Tests of Model Effects			
Parameter	Wald Chi-Square	df	Sig.
(Intercept)	379.893	1	<.001
PMI	0.88	2	0.644
DIS	34.666	1	<.001
PL	67.382	2	<.001
DM	66.948	2	<.001

Table 8. Wald chi-square test indicating strength of properties to predict ratings of idiomaticity.

4.2.1.1.1. The intercept

In Tables 6-8, the intercept represents the predicted value of the dependent variable when the categoric independent variables are at their reference level. The selected reference levels correspond to the level hypothesized to be the least prototypically idiomatic and, therefore, the least predictive of perceived idiomaticity. The reference levels were: +DIS (a phrase has distributed meaning) for distributed idiomatic meaning, -PL_TwoLit (not partially literal, both words contribute a meaning they convey outside of the phrase) for partial literality, -DM_LitOnly (no dual idiomatic meaning, a phrase has a plausible literal interpretation only) for dual meaning, and PMI-L (low PMI) for formulaicity. The intercept was significant ($\beta = -1.92$, $z = 19.49$, $p < .001$), indicating that the model predicts ratings of idiomaticity at a statistically significant level when the reference levels for the tested properties are set at their hypothesized least prototypical values. This means that, while not all individual properties independently predict idiomaticity, the model as a whole captures meaningful variance, demonstrating the importance of the included properties for explaining idiomaticity ratings.

4.2.1.1.2. Distributed idiomatic meaning

Distributed idiomatic meaning was a significant predictor of perceived idiomaticity. Phrases with no distributed meaning (-DIS, ß=0.397, z=5.831, p<.001) were significantly more likely to be rated as idiomatic than phrases with distributed idiomatic meaning, which served as the reference level (+DIS). This finding highlights the degree to which phrases with no distributed meaning are likely to be rated as idiomatic, supporting a continuous account of the class idiom in which phrases that do not have distributed idiomatic meaning (e.g., *sweet tooth*) are more likely to be rated as idiomatic compared to phrases with distributed idiomatic meaning (e.g. *big head*). Further, it suggests that, at the construct level, distributed idiomatic meaning is a distinguishing characteristic of idiomaticity as the model shows that distributed idiomatic meaning is a significant predictor of perceived idiomaticity.

The individual contribution strength of distributed idiomatic meaning was assessed using a Wald chi-squared test. Given the degrees of freedom, a Wald value of 3.84 was the threshold to establish significance for distributed idiomatic meaning at a p=.05 level. The observed Wald value for distributed idiomatic meaning was Wald $\chi^2(1) = 34.666$ (see Table 9). This is significant at the level of p<.001, indicating that distributed idiomatic meaning is a strong predictor of idiomaticity and makes a significant contribution to the model.

Parameter	Wald Chi-Square	df	Sig.
DIS	34.666	1	<.001
+DIS	Reference	.	.
-DIS	33.999	1	<.001

Table 9. Wald chi-squared estimates for distributed idiomatic meaning.

No significant correlations were observed between distributed idiomatic meaning and any other variables. The lack of direct correlations with other variables suggests that distributed idiomatic meaning may function independently in predicting idiomaticity. This means that changes in distributed idiomatic meaning do not necessarily coincide with changes in another. This independence aligns with theoretical claims suggesting that distributed idiomatic meaning reflects a unique cognitive construct associated with idiomatic phrases, reinforcing its status as a prototypical feature of idioms. While future work should consider the possibility of interactions between uncorrelated variables, this finding supports claims that distributed idiomatic meaning is an independent construct.

4.2.1.1.3. Partial literality

Partial literality was a significant predictor of perceived idiomaticity. To rate partial literality, participants determined whether both (-PL_TwoLit), one (+PL_OneLit), or neither (-PL_NoLit) of the words in a phrase contributed a meaning they can independently convey outside of a test phrase. While this analysis considered ratings of partial literality for dictionary idioms only, it is important to note that participants were free to indicate that, upon reflection, they believed both words in an idiomatic phrase were used literally (-PL_TwoLit). While one might not expect this response option to be selected for idioms, it was selected by participants at least occasionally[25]. An item analysis showed that these phrases corresponded to dictionary idioms theorized to be perceived as weakly idiomatic.

Phrases rated as partially literal (+PL_OneLit) were significantly more likely to be rated as idiomatic (β=0.584, z=7.509, p<.001) than phrases in which both words

[25] Selection of -PL_TwoLit appears to reflect participants' interpretations rather than an issue with instrument validity. In a pilot study where participants rated idiomaticity immediately after partial literality, participants still identified some phrases as idiomatic even when they rated both words as literal. This pattern suggests that the -PL_TwoLit choice was a legitimate reflection of participant perceptions rather than a methodological artifact.

contributed a meaning they convey outside of the phrase (-PL_TwoLit), which served as the reference level. Similarly, phrases rated as not partially literal because neither word conveyed a meaning it has outside the phrase (-PL_TwoLit) were significantly more likely to be rated as idiomatic than the reference level (+PL_NoLit, ß= 0.633, z = 6.257, p < .001). Thus, phrases that are not partially literal because both words convey a meaning they can have outside of the phrase are less predictive of idiomaticity compared to the other levels. To determine whether phrases that were not partially literal because neither word conveys a meaning that it can have outside of the phrase (-PL_NoLit) was significantly more predictive of idiomaticity than +PL_OneLit, a log-odds analysis and pairwise comparison were conducted. The log-odds for phrases in which one word contributes a non-idiomatic meaning (+PL_OneLit) were 0.049 lower than those for phrases in which neither word contributes a literal meaning (-PL_NoLit). This indicates that partially literal phrases were slightly less likely to be perceived as idiomatic than those with no literal contributions. However, this difference was minimal and not statistically significant (p = 0.83).

The use of a word or words in a non-literal manner often makes them stand out as unusual because their individual meanings may stand in stark contrast to the kinds of words that would be expected in a given context. When both words in a phrase are used in an unusual manner, or a way in which they are not used when encountered on their own outside of the phrase, the phrase stands out more than if only one word were used in an unusual manner. Additionally, when a phrase is partially literal, the word conveying the same meaning within the phrase as outside of it often establishes a reference domain that may make it more difficult to determine whether the other word is used idiomatically or whether the phrase has been metaphorically extended. For example, while the phrase *laundry list* does not

refer to an actual "list about laundry," it does refer to a type of list, often in a household context. *Laundry* establishes a reference domain situated within the conceptual background of household chores or repetitive tasks and indirectly indexes it by evoking associations with chores. This indexing function supports a metonymic reading where *laundry* stands in for a broader set of household responsibilities. Because *laundry list* retains a tangible connection to this literal domain, it may be perceived as less idiomatic as its referential and indexical properties keep it grounded in concrete, known context.

The individual contribution strength of partial literality was assessed using Wald chi-squared. At the construct level, the observed Wald value for partial literality was Wald $\chi^2(1) = 67.382$, see Table 10). This is significant at the level of $p<.001$, indicating that partial literality makes a significant contribution to the model and can predict which phrases will be reliably rated as idiomatic. Within the construct, phrases rated as not partially literal because neither word conveyed a meaning they can have outside of the phrase (-PL_NoLit] made a significant contribution to the model (Wald $\chi^2(1) = 39.144$, $p<.001$). The contribution of partially literal phrases (+PL_OneLit) was even stronger (Wald $\chi^2(1) = 56.388$, $p<.001$). This indicates that while both partially literal phrases (+PL_OneLit) and phrases that are not partially literal because neither word conveys a meaning it has outside of the phrase (-PL_NoLit) significantly contribute to the prediction of idiomaticity ratings, partial literality (+PL_OneLit) has a stronger impact on the model. Despite the finding that phrases in which neither word is used literally (-PL_NoLit) is more strongly associated with idiomaticity than partial literality (+PL_OneLit), the variability in idiomaticity ratings explained by partial literality (+PL_OneLit) is greater. In other words, partial literality (+PL_OneLit) plays a more significant role in the model's ability to distinguish between idiomatic and non-idiomatic phrases,

even though phrases in which neither word conveys a meaning it has outside of the phrase (-PL_NoLit] has a higher effect size in terms of its association with idiomaticity.

Parameter	Wald Chi-Square	df	Sig.
PL	67.382	2	<.001
-PL_TwoLit	Reference	.	.
+PL_OneLit	56.388	2	<.001
-PL_NoLit	39.144	2	<.001

Table 10. Wald chi-squared estimates for partial literality.

No correlations were observed between partial literality and any other variable, suggesting that changes in one variable does not necessarily coincide with changes in the relationship between partial literality and idiomaticity. This supports claims that partial literality is an independent construct. Overall, the findings indicate that partial literality is a strong predictor of idiomaticity, with each level making a significant contribution. This suggests that partial literality is associated with the cognitive construct of idiomaticity and may serve as a prototypical feature of the category.

4.2.1.1.4. Dual meaning

Plausible dual literal and idiomatic meaning was a significant predictor of perceived idiomaticity. To rate dual meaning, participants determined whether a phrase had a plausible literal interpretation only (-DM_LitOnly), a plausible literal and non-literal interpretation (+DM_Both), or a plausible non-literal interpretation only (-DM_IdOnly).

In this analysis, phrases rated as having a plausible literal interpretation only (-DM_LitOnly) served as the reference level. Like -PL_TwoLit, one might not expect this option to be selected for a dictionary idiom because it seems fundamentally incompatible. However, many participants commented that phrases that "felt" idiomatic also "seemed to have more than one literal meaning but no clearly idiomatic meaning." Because participants were asked to think of various contextually driven meanings that a phrase could have rather than providing a judgement of whether a phrase in a given context was idiomatic or non-idiomatic, this option was included in the idiom analysis.

Phrases rated as having both plausible literal and non-literal interpretations (+DM_Both) were significantly more likely to be rated as idiomatic compared to phrases with only a plausible literal meaning (-DM_LitOnly, ß=0.244, z=3.227, p=.001). Similarly, phrases with no plausible literal meaning (-DM_IdOnly) were significantly more likely to be rated as idiomatic compared to phrases with only a plausible literal meaning (-DM_LitOnly, ß=0.769, z=7.951, p<.001). Thus, these two levels are more significant predictors of idiomaticity than the reference level.

To explore this further, a log-odds analysis and pairwise comparison were conducted to determine whether phrases with no plausible literal meaning were more predictive of idiomaticity than phrases with both plausible literal and non-literal meanings. The log-odds for phrases with both meanings were 0.525 lower than for phrases with no plausible literal meaning, indicating that phrases with an exclusively idiomatic meaning are more likely to be perceived as idiomatic. This difference was statistically significant (p=.041). This finding supports the prediction that phrases with no plausible literal meaning are stronger predictors of perceived idiomaticity than those with both plausible literal and non-literal interpretations.

When a phrase has no plausible literal interpretation, it stands out, making it more easily recognizable as idiomatic. By contrast, phrases with both literal and non-literal interpretations generally stand out less, particularly if there is a logical relationship or mapping between the non-idiomatic and idiomatic interpretations. For example, the idiomatic interpretation of *big picture* is a metaphoric extension of the plausible non-idiomatic interpretation. This result aligns with cognitive theories suggesting that phrases with only a non-literal interpretation have reduced processing demands, allowing for easier, more reliable and consistent judgments of idiomaticity (cf. Gibbs et al. 1989, Gibbs et al. 1997, Libben & Titone 2014, Morid 2021).

To assess the individual contribution strength of dual meaning, a Wald chi-squared test was conducted. At the construct level, the observed Wald value for dual meaning was (Wald $\chi^2(1) = 66.948$, see Table 11), which is significant at the level of $p<.001$, indicating that dual meaning makes a significant contribution to the model and can predict which phrases will be reliably rated as idiomatic. Within the construct, dual meaning (+DM_Both) made a significant contribution to the model (Wald $\chi^2(1) = 10.41$, $p=.001$). The contribution of phrases rated as having a non-literal meaning only (-DM_IdOnly) was also significant (Wald $\chi^2(1) = 63.217$, $p<.001$). Thus, dual meaning and non-literal meaning only are significantly predictive of idiomaticity, with phrases that have a non-literal meaning only (-DM_IdOnly) more strongly impacting the model.

Parameter	Wald Chi-Square	df	Sig.
DM	66.948	2	<.001
-DM_LitOnly	Reference	.	.
+DM_Both	10.41	1	0.001
-DM_IdOnly	63. 217	1	<.001

Table 11. Wald chi-squared estimates for dual meaning.

No correlations were found between dual meaning and other variables, supporting the claim that dual meaning is a fully independent construct. This suggests that, like distributed idiomatic meaning and partial literality, dual meaning is a distinct factor in idiom conceptualization and comprehension. This also implies that controlling for other variables alone may not fully account for the significant role that dual meaning plays in these processes.

Overall, the findings indicate that dual meaning is a strong predictor of idiomaticity, with each level making a significant contribution. This suggests that dual meaning is associated with the cognitive construct of idiomaticity and may serve as a prototypical feature of the category.

4.2.1.1.5. Formulaicity

Formulaicity was not a significant predictor of perceived idiomaticity. As in experiment 1, three levels of PMI were included in the analysis, high PMI (PMI-H), which had a PMI value of PMI>6 and represented a high degree of formulaicity, medium PMI (PMI-M), which had a PMI value between 5.9<PMI>3, and low PMI (PMI-L), which had a PMI value of PMI<2.9. Highly formulaic phrases were less likely to be rated as idiomatic than those with a low degree of formulaicity, but this difference did not reach significance (ß=-.057, z=.322, p=.401). Phrases with a medium degree of formulaicity were more likely to be rated as idiomatic than those

with a low degree of formulaicity, but this difference also failed to reach significance (ß=.049, z=.597, p=.551).

To further evaluate formulaicity, the contribution strength of the construct as well as each PMI level was assessed. At the construct level, the observed Wald value for formulaicity was (Wald $\chi^2(1) = .88$, p=.644, see Table 12), indicating that, overall, formulaicity does not make a significant contribution to the model. Within the construct, the contribution of medium formulaicity was (Wald $\chi^2(1) = .356$, p=.551) and the contribution of high formulaicity was (Wald $\chi^2(1) = -.103$, p=.401). Additionally, no correlations were observed between formulaicity and any other variables, suggesting that formulaicity is an independent construct. Thus, the lack of significance cannot be attributed to complimentary or shared variation in another variable.

Parameter	Wald Chi-Square	df	Sig.
PMI	0.88	2	0.644
PMI-L	Reference	.	.
PMI-M	0.356	1	0.551
PMI-H	-0.103	1	0.401

Table 12. Wald chi-squared estimates for formulaicity.

Taken together, these findings indicate that formulaicity is not a significant predictor of perceived idiomaticity and suggest that PMI does not significantly contribute to the model's ability to accurately predict ratings of idiomaticity. This finding may be surprising at first as one might expect formulaicity to be a distinguishing feature of idioms due to the fact that the more formulaic a phrase is, the more conventionalized it becomes, and the fact that the degree of

conventionality is positively associated with semantic narrowing (Sanchez Lopez 2015). However, there is no straightforward relationship between meaning narrowing and idiomaticity as narrowing does not necessarily make a phrase more idiomatic. For example, the phrases *ice cube, fruit bowl,* and *center divider* are highly formulaic. However, few would argue that they are idiomatic as they denote literal, physical objects. For semantic narrowing to result in idiomaticity, the phrase must not only establish a more specific referent as canonical, there must also be motivation sufficient to further narrow the meaning to one that overrides the individual, non-idiomatic meaning of the individual constituents, promoting an idiomatic interpretation. As noted by Nunberg et al. (1994:493), this is most likely to occur in socially charged situations, particularly those that imply an evaluative stance or directly naming difficult topics (e.g., *kick the bucket* rather than *die*) and in recurrent social situations (e.g., *spilling beams, breaking ice*). This analysis is supported by the findings, which suggest that, while all idioms are formulaic, formulaicity is not a distinguishing feature of idiomaticity. Instead, this property may be reflective of collocations more generally.

4.2.1.1.6. Properties as indicators of prototypical idiomaticity

The results of the regression analysis highlight significant relationships between distributed idiomatic meaning, partial literality, dual meaning, and idiomaticity ratings, indicating that these properties impact ratings of idiomaticity.

Overall, the findings support the predictions. First, the results support the prediction that distributed idiomatic meaning would be a significant predictor of perceived idiomaticity and suggest that distributed idiomatic meaning is a distinguishing feature of idiomaticity. The findings further supported the prediction that phrases with no distributed idiomatic meaning would be recognized as idiomatic more often than those with distributed idiomatic meaning, suggesting

that phrases with no distributed idiomatic meaning (e.g., *sweet tooth, sack time*) are perceived as more prototypically idiomatic than phrases with distributed idiomatic meaning (e.g., *big picture, slim chance*). This finding is in line with prior work showing that the more analyzable an idiomatic phrase is, the more it is treated as though it were non-idiomatic (Gibbs & Nayak 1989, Hamelin & Gibbs 1999). In light of this, I suggest that phrases that are highly analyzable are likely to be perceived as less idiomatic than those that are non-analyzable because the ability to decompose an idiom and understand it in terms of its parts makes it appear less like a single, holistic unit.

Further support for this claim comes from work on idiom comprehension, which has shown that phrases with distributed idiomatic meaning (analyzable phrases) are processed more quickly than those with no distributed idiomatic meaning because individual constituents can be quickly parsed and combined to derive the overall figurative meaning (Gibbs 1993, Gibbs et al. 1989, Hamelin & Gibbs 1999, Tabossi et al. 2010). Conversely, phrases with no distributed idiomatic meaning may require holistic meaning retrieval, which may occur only after an attempt at an individual constituent analysis fails (cf. Weinreich 1969, Bobrow & Bell 1973).

The results also support the prediction that partial literality would be a distinguishing feature of idioms and that phrases in which no words convey a meaning inside the phrase that they can have outside of the phrase would be a stronger predictor of perceived idiomaticity, with these phrases more likely to be rated as idiomatic. Like distributed idiomatic meaning, this can be partially explained in terms of analyzability. While distributed idiomatic meaning and partial literality are independent - a partially literal phrase may or may not have distributed idiomatic meaning - partially literal phrases are viewed as at least partially analyzable due to overlapping meaning for one word. This overlap may

also add a degree of transparency. Further, while there is no relationship between partial literality and flexibility, partial literality creates an illusion of reduced fixedness. With respect to comprehension, the overlapping meaning activation that occurs in partially literal phrases provides two sources of activation, one literal and one idiomatic. While this could delay the point at which one commits to an idiomatic interpretation over a non-idiomatic interpretation, the dual activation is thought to facilitate comprehension since there is no competition between the meanings (cf. Gibbs et al. 1989, Titone & Connine 1999, Libben & Titone 2008). Dual activation may also strengthen the association between the idiomatic and non-idiomatic representations, resulting in less clear categoric boundaries between phrases.

With respect to dual meaning, the results support the prediction that this property would serve as a significant predictor of idiomaticity. Further they support the prediction that phrases with no dual meaning are more likely to be rated as idiomatic than those with dual meaning. When a phrase has dual meaning, often, there is a motivated relationship between the idiomatic and non-idiomatic interpretations of a phrase. When this is true, the line between an idiomatic interpretation and a non-idiomatic interpretation is blurred. Phrases used in this experiment fell into this common grouping, including phrases such as *big picture, track record, laundry list, moving target,* and *blank slate,* which have transparent relationships between the non-idiomatic and idiomatic interpretations. This fact further motivated the prediction that phrases with dual meaning would be less predictive of idiomaticity than phrases with no dual meaning.

However, dual meaning should be used with caution as there is not always a transparent relationship between interpretations. The impact of this difference can be seen in comprehension work, such as that by Beck & Weber (2020), who found a facilitation effect for phrases with transparent dual meaning but observed no

facilitation effect for phrases with no dual meaning. Crucially, dual meaning phrases with an opaque relationship between the idiomatic meaning and non-idiomatic meaning were associated with slowed processing.

If the idiomatic interpretation of a phrase with dual meaning is not transparently related to its non-idiomatic interpretation, it should be easier to recognize because the non-literal meaning of these phrases stands in direct contrast to the literal meaning. This juxtaposition may cause the idiomatic meaning to stand out more so than for phrases with no dual meaning. Over time, the "weirdness" initially associated with a new-to-you idiomatic phrase diminishes. Such habituation may cause a phrase with no dual meaning to stand out comparatively less, facilitating processing and diminishing the degree to which dual meaning independently contributes to perceived idiomaticity.Conversely, direct comparison with the simultaneously activated-literal interpretation that occurs for phrases with opaque dual meaning may mitigate habituation to the unexpected idiomatic meaning to some degree. As a result, there may be a positive correlation between familiarity and recognition of phrases with opaque dual meaning, such that the speed and consistency by which opaque dual meaning phrases are recognized as idiomatic increases as familiarity increases. Thus, while the results indicate that phrases with no plausible dual literal and idiomatic meaning are more likely to be deemed idiomatic as competing interpretations reinforce the sense that the phrase cannot be understood through a fully literal lens, a more nuanced analysis may reveal this to be an oversimplification. Instead, it may suggest a difference between phrases with transparent versus opaque dual meanings such that, when separately considered, phrases with opaque dual meaning are more reliably recognized as idiomatic. Findings supporting these predictions would be indicative of scalar

judgments, supporting a prototype-based approach and adding weight to the notion that phrases exist along a continuum of idiomaticity.

Finally, the results support the prediction that formulaicity would not be a significant predictor of idiom conceptualization based on analytic tasks. While formulaicity significantly impacts comprehension speed, it does not impact idiom conceptualization such that it can be relied upon to aid in the differentiation of idioms from non-idioms at least in tasks relying on metalinguistic, analytic methodologies. This finding shares some similarities with those of Titone & Connine (1994b), who considered the role of word co-occurrence frequency, familiarity, and transparency. In an untimed acceptability judgment task, transparency and familiarity impacted ratings. However, while frequency effects were seen in reaction time experiments, no effect was observed in the ratings tasks, which allowed ample time for participants to consider a phrase, far exceeding the level of attention and effort expended during normal processing (e.g., everyday conversation). The reaction time finding was attributed to entrenchment. When a phrase is highly formulaic, there is no need to recompute its meaning every time it is encountered. Instead, as a phrase becomes more predictable, the association strength between its constituents increases. During comprehension, as a sentence progresses, the surprisal rate for upcoming words tends to decrease. When a sentence ends with a highly formulaic phrase, the surprisal rate of upcoming words decreases exponentially as predictors come from two sources. The high degree of association allows for quick retrieval of collocates at a time point prior to or early in the window associated with semantic analysis (c.f. Vespignani et al. 2010, Rommers et al. 2013, Canal et al. 2017). Importantly, this process is not specific to idioms. The phrases *fruit bowl, ice cube, bad apple,* and *sweet tooth* are all highly formulaic and are likely to be processed more quickly than phrases with a low degree of

formulaicity. Additionally, findings from Carrol & Conklin (2019) suggest that there should be no significant difference in the processing time required for the non-idiomatic *fruit bowl* or *ice cube* versus the idiomatic *bad apple* or *sweet tooth*, despite the fact that these phrases differ in the properties they express. Libben & Titone (2008) explain this in terms of a *masking effect* of formulaicity, noting that formulaicity may impact comprehension at an earlier time point within the processing window than other properties. When a phrase is highly formulaic, the associative strength between initial words allows for quick retrieval of upcoming words. Retrieval does not depend on performance of a full semantic analysis, and findings indicate that, under certain conditions, retrieval may even be completed before the time point at which a full analysis would be expected to begin (Libben & Titone 2008, Carrol & Conklin 2019). This is not to say that the phrase is holistically represented or that properties do not impact comprehension. Semantic analysis continues for predicted words and can be seen when a task takes property masking into account by factoring out or controlling for the early effects of formulaicity and by considering multiple time points during idiom comprehension (c.f. Libben & Titone 2008, Carrol & Conklin 2019).

4.2.1.2. A continuum of prototypical idiomaticity

Based on the results of experiment 2, we can conclude that distributed idiomatic meaning, partial literality, and dual meaning describe the mental construct of idiomaticity. But, do they serve as markers of prototypicality? While the regression establishes their influence, it does not directly indicate the distribution or frequency of ratings across different property combinations (e.g., +DIS, +PL_Both, +DM_Both or -DIS, -PL_NoLit, -DM-NoLit). If properties are associated with prototypical idiomaticity, then we should be able to model a continuum of idiomaticity based on the relationships identified in the regression. This was investigated using a series of

t-tests to compare the differences in idiomaticity ratings for phrases corresponding to each of the possible property combinations (e.g., +DIS, +PL_Both, +DM_Both, -DIS, +PL_Both, +DM_Both, +DIS, -PL_IdOnly, +DM_Both, etc.).

These tests revealed a continuum of perceived idiomaticity (see Figure 6). At one end of the continuum are phrases with no distributed idiomatic meaning, that are not partially literal, and that do not have dual meaning (-DIS, -PL, -DM, e.g., *sweet tooth*). These phrases were rated as idiomatic significantly more often than any other property combination ($p<.001$). At the other end of the continuum are phrases with distributed idiomatic meaning, that are partially literal and that have dual meaning (+DIS, +PL, +DM, e.g., *big head*). These phrases were rated as idiomatic significantly less often than any other property combination ($p<.001$). In the middle are the other potential property combinations. For example, the distributional frequencies indicate that -DIS, -PL, +DM (e.g., *butcher block*) is more highly associated with prototypical idiomaticity than +DIS, -PL, +DM (e.g., *blank slate*), which is more associated with prototypical idiomaticity than +DIS, +PL, -DM phrases (e.g., *slim chance*). Thus, not only do properties impact perceived idiomaticity but the distribution of ratings across different property combinations suggests that distributed idiomatic meaning, partial literality, and dual meaning are associated with prototypical idiomaticity.

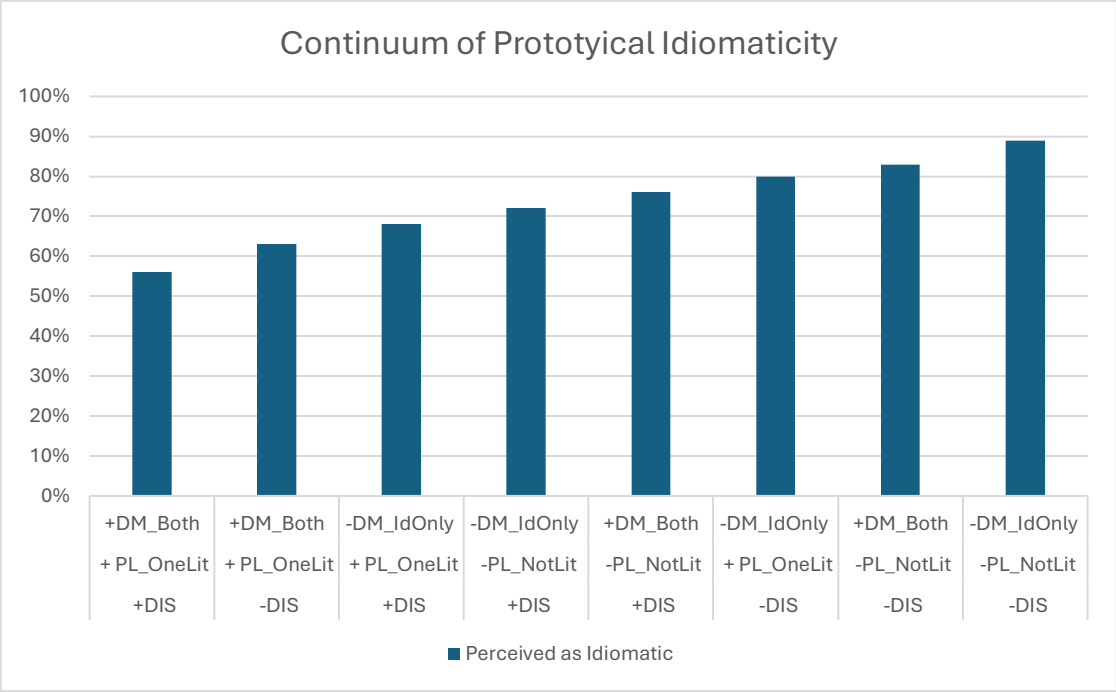


Figure 6. Continuum of idiomaticity based on collected ratings.

In conclusion, the results of experiment 2 support the prediction that certain combinations of properties make certain idioms stand out as idiomatic more so than others. This finding is in line with the findings of experiment 1, adding weight to the notion that phrases exist along a continuum of idiomaticity and supporting a prototype-based approach in which properties are not only related to idiomaticity but partially define it such that the presence of more properties is associated with a higher degree of perceived idiomaticity while the presence of fewer properties is associated with a lower degree of perceived idiomaticity.

CHAPTER V

GENERAL DISCUSSION AND CONCLUSION

The goal of this work was to address gaps in psycholinguistic and linguistic research regarding the mental construct of idiomaticity. This was achieved by investigating factors that influence idiom perception and conceptualization, specifically addressing the challenge of defining the cognitive construct of idiomaticity and identifying what makes a phrase recognizable as an idiom. The research was motivated by two central questions: first, whether the mental construct of idiomaticity is dichotomous or continuous, and second, the role of idiom properties in determining the degree of prototypicality associated with a given phrase.

Specifically, RQ1a asked: Are native speakers able to differentiate between idioms and non-idiomatic collocations? Building on this, RQ1b refined the investigation by asking: Do dictionary-defined idioms receive higher ratings of idiomaticity from native speakers than equally formulaic non-idiomatic collocations? These questions addressed the assumption that idioms form a cleanly delineated mental category about which native speakers share intuitions. They also questioned whether noncompositionality, which is often used synonymously with idiomaticity within psycholinguistics, is sufficient as the single defining criterion to differentiate idioms from non-idiomatic phrases, thereby accounting for perceived idiomaticity.

Following this, RQ2 asked: Are formulaicity, distributed idiomatic meaning, partial literality, and dual meaning able to predict which phrases will be reliably recognized as idiomatic? This question tested assumed relationships between four

idiom properties and the cognitive construct of idiomaticity. It sought to determine whether these properties directly reflect the conceptualization of a phrase as a member of the idiom class.

Overall, the findings suggest that idiomaticity is not a discrete, binary construct but exists on a continuum, with some idioms perceived as more prototypical than others. The degree of prototypicality can be predicted by considering certain combinations of properties exhibited by a phrase. These findings challenge a single criterion definition of idiomaticity, which assumes that non-compositionality is sufficient to differentiate idiomatic phrases from non-idiomatic collocations. Instead, they support a multi-determinate account of idiomaticity in which a number of properties impact how we think about and understand idioms. Thus, the findings suggest that these properties are an ideal place to begin to decompose cognitive idiomaticity and this research lays the groundwork for future studies considering additional properties that impact idiomaticity allowing for further refinement of the category idiom.

5.1. Experiment 1

Experiment 1 tested assumptions about the class idiom, seeking to determine whether the mental category is discrete or whether a continuous approach is more appropriate by considering the degree to which native speakers share intuitions about class members. This was accomplished by collecting idiomaticity judgments for idioms and equally formulaic, non-idiomatic collocations. The results revealed that there is no categorical distinction between idioms and non-idioms. Native speakers of American English were not able to reliably differentiate idioms from equally formulaic non-idiomatic collocations. The findings of experiment 1 suggests that idiomaticity is a graded construct and support an alternative definition of the class idiom. While some phrases were reliably differentiated as idiomatic or non-

idiomatic, others were less reliably rated. At times, dictionary idioms were deemed less idiomatic than non-idioms. However, a general trend was observed, revealing that, dictionary idioms were rated as idiomatic significantly more often than non-idioms overall. Together, these findings support the existence of a cognitive class idiom and suggest that idiomaticity is a graded construct.

These findings have significant implications for both psycholinguistic theory and research methodology. Research considering how we think about and understand idioms has often relied on the assumption that native speakers intuitively recognize idioms as distinct from other types of phrases, generally using this assumption to structure experiments and interpret findings. However, the results suggest that such experiments may be operating under a false premise. The results suggest that native speaker intuitions about idioms may not be as consistent as previously thought, especially when it comes to differentiating idioms from non-idiomatic phrases. This inconsistency challenges the assumption that idioms form a distinct, universally recognized category in the mental lexicon. Instead, it appears that idiomaticity is influenced by a range of properties that impact how phrases are conceptualized and understood. If idioms are not easily distinguishable from non-idiomatic phrases, then experimental results that depend on this distinction may be flawed or incomplete.

Moreover, the lack of reliable differentiation highlights the need for more precise norming and validation of idiom lists in experimental work. Researchers cannot assume that dictionary-defined idioms will always be perceived as idiomatic by participants, particularly when formulaicity is held constant. Future studies will need to employ more rigorous methodologies for testing idioms and non-idiomatic phrases to ensure that they are truly investigating the cognitive processes specific to idiomaticity, rather than formulaicity or other confounding factors.

Finally, the implications extend beyond experimental design. The assumption that idioms form a discrete category is central to many models of idiom comprehension, particularly holistic models that posit a specialized processing strategy for idioms. The findings challenge these models by showing that native speakers' intuitions about idioms are not as clear-cut as these models assume. Instead, a more nuanced understanding of idiomaticity is needed, one that accounts for the graded nature of idiomaticity and the variation in how phrases are perceived. This motivated experiment 2, which analyzed the relationships between perceived idiomaticity and four idiom properties, establishing a prototype-based description of the cognitive class idiom.

5.2. Experiment 2

In experiment 2, ratings of distributed idiomatic meaning, partial literality, and dual meaning were collected. These ratings, along with previously calculated formulaicity scores (PMI), were used to evaluate relationships between idiom properties and perceived idiomaticity.

The results suggest that three of the four tested properties are defining features of the mental category idiom and that they significantly influence how native speakers categorize phrases. First, distributed idiomatic meaning emerged as a key factor in distinguishing idiomatic phrases. Phrases with distributed idiomatic meaning, or phrases in which individual words each contribute meaning to the idiom, are more likely to be perceived as idiomatic compared to phrases with no distributed idiomatic meaning. This suggests that the analyzability of an idiom plays a role in its perceived idiomaticity, indicating that idioms are not purely holistically processed, as some models suggest. Future work should also consider the role of flexibility. Gibbs & Nayak (1989) found a positive correlation between distributed idiomatic meaning and flexibility such that phrases with distributed idiomatic

meaning were more flexible than those with no distributed idiomatic meaning. Knowledge of how individual words combine allow speakers to use phrases creatively. When a phrase has no distributed idiomatic meaning, it is less likely to undergo certain creative processes such as lexical substitution, internal modification, or constituent movement, because the exact portion of the phrase one intends to modify cannot be isolated. Phrases with distributed idiomatic meaning are more likely to allow such processes as the modified constituent can be isolated. This type of analyzability, which allows for creative, flexible use, is more associated with non-idiomatic phrases than prototypically idiomatic phrases.

Partial literality also significantly impacted idiom recognition. Phrases in which neither word contributes a non-idiomatic meaning are more likely to be perceived as idiomatic than phrases in which one word is used idiomatically and the other is used non-idiomatically or those in which both words are used non-idiomatically. Unlike phrases with distributed idiomatic meaning, partially literal phrases may not be fully analyzable. For example, a *pipe dream* is a type of "dream", but no meaning can be assigned to *pipe.* Instead, partially literal phrases are differently analyzable in that the relationship between the word conveying a non-idiomatic meaning establishes a literal referent, generally creating a transparent connection useful in decoding the meaning of the phrase. This finding is in line with theories positing active evaluation of both idiomatic and non-idiomatic meanings during processing. The dual activation of literal and idiomatic meanings may facilitate quicker recognition of idiomatic phrases.

Dual meaning (DM), which refers to whether a phrase has a plausible idiomatic and non-idiomatic interpretation or whether it can be used idiomatically only, was also a significant predictor. Phrases with no dual meaning are more likely to be perceived as more idiomatic than those with an idiomatic and non-idiomatic interpretation,

suggesting that the lack of a plausible non-idiomatic interpretation causes a phrase to stand out as idiomatic more so than a phrase with plausible idiomatic and non-idiomatic interpretations, particularly if there is a logical relationship or mapping between the non-idiomatic and idiomatic interpretation (for further discussion, see section 5.4).

Collectively, properties can be used to create a continuum of perceived idiomaticity. At one end of the continuum are phrases such as *moving target* and *big picture* that have distributed idiomatic meaning, are partially literal, and have dual meaning. These phrases are perceived as the least idiomatic. At the other end of the continuum are phrases such as *sweet tooth* and *funny bone* that have no distributed idiomatic meaning, are not partially literal, and have no dual meaning. These phrases are perceived as the most idiomatic. In the middle are phrases such as *seat warmer, lead foot, fresh mouth,* and *tight spot.* Phrases like *seat warmer* and *lead foot,* which have distributed idiomatic meaning, are not partially literal, and have dual meaning, are perceived as relatively more prototypically idiomatic than phrases such as *fresh mouth* and *tight spot,* which have distributed idiomatic meaning, are not partially literal, and do not have dual meaning, as they exhibit one more prototypical feature, that of dual meaning. The finding that each combination of properties was predictive of idiomaticity to a different degree further supports a prototype-based approach. Further, the degree to which a property combination predicts perceived idiomaticity does not always align with the predictive contributions of individual properties. This has methodological implications for future research, as it suggests the presence of more complex relationships and highlights the potential for experimental findings to be affected when all properties are not considered. For example, even though phrases with dual meaning are perceived as more idiomatic than those with no dual meaning, phrases

with dual meaning, that are partially literal, and that have distributed idiomatic meaning (e.g., *moving target, big picture, track record*) are perceived as the least idiomatic of the property clusters. Future work should consider whether properties interact with each other, impacting their relationships with perceived idiomaticity.

One property, formulaicity, was not a significant predictor of perceived idiomaticity. This does not mean that formulaicity is unrelated to idiomaticity as all idioms are formulaic. Instead, it reflects the important fact that idioms are not the only type of formulaic language. Thus, this finding should be interpreted as indicating that formulaicity is not a distinguishing feature of idiomaticity in that it cannot be used to differentiate an idiom from a non-idiom. Several considerations support this conclusion.

First, there is no straightforward relationship between the degree of formulaicity and how "weird" or unexpected a phrase is. The collection of properties associated with a high degree of prototypicality can be summarized as those that cause the phrase to stand out as idiomatic, even after frequent exposure results in habituation to the conventions surrounding its use. Phrases with high prototypicality[26] stand out more than less prototypical idioms[27], regardless of their level of formulaicity. In contrast, phrases that are highly formulaic but lack prototypical features may not be perceived as strongly idiomatic. In other words, highly prototypical phrases with a low degree of formulaicity will be perceived as more idiomatic than less prototypically idiomatic phrases with a high degree of formulaicity. According to Gibbs (1995:61), a high degree of formulaicity should be associated with increased anticipation of idiomatic meaning during online comprehension. This prediction was upheld by Carrol & Conklin (2019) who showed

[26] Phrases with no distributed idiomatic meaning, that are not partially literal, and that have no dual meaning.
[27] Phrases with distribute idiomatic meaning, that are partially literal, and that have dual meaning.

that familiar, highly formulaic phrases were comprehended more quickly than less formulaic phrases. This was true for non-idiomatic collocations as well; phrases with a high degree of formulaicity were comprehended more quickly than those with a low degree of formulaicity. While this suggests that formulaicity plays an important role in idiom comprehension, it indicates that the role played by formulaicity is not unique to idioms but is seen with other types of conventionalized, non-idiomatic phrases. Because formulaicity is not unique to idioms and its impact has been documented with other types of language, there is no reason to assume that formulaicity could serve as a differentiating feature of the class idiom.

While no effect of formulaicity on idiom conceptualization was observed, this finding deserves further consideration. The role of formulaicity in online comprehension has been established and, such work shows that formulaicity impacts comprehension at a very early time point prior to semantic and syntactic analysis. In light of this, one might not expect to see an impact of formulaicity in an analytic task, which allows ample time for subconscious processing and conscious analytic reflection. It is also likely that no effect was observed because formulaicity is not a defining feature of idiomaticity. While all idioms are, by definition, formulaic, they are not the only type of formulaic language (Wray & Perkins 2000, Wray 2012, see also Howarth 1998).

Overall, these findings provide important insight into the mental construct of idiomaticity, demonstrating that distributed idiomatic meaning, partial literality, and dual meaning moderate perceived idiomaticity, serving as indicators of prototypicality that can be used to predict the likelihood that a phrase will be rated as idiomatic. This suggests that idioms are not processed as single lexical units (words_with_spaces), but rather that their internal structure plays a crucial role in

how they are conceptualized, with each property contributing to their mental representation.

The results of experiment 2 highlight the complexity of idiom processing and challenge the notion that idiomaticity can be explained by a single property, such as noncompositionality. Instead, the results reinforce the notion that idiomaticity is a graded, multidimensional construct influenced by multiple properties. The fact that no single property accounted for idiomaticity across all phrases highlights the complexity of this construct and underscores the importance of a continuous approach to account for idiomaticity. This is further supported by the individually and collectively significant roles played by distributed idiomatic meaning, partial literality, and dual meaning, which suggest a prototype-based approach may be appropriate as different phrases exhibited varying degrees of idiomaticity based on the properties they display. Idioms that exhibit certain properties to a greater degree may be more prototypical members of the idiom class, while those that lack these certain properties may be closer to non-idiomatic collocations. Within linguistics, a prototype-based account aligns with theories of language that emphasize the role of gradience and variation, rather than adopting rigid categorical boundaries. This has important implications for models of idiom conceptualization and comprehension. Rather than relying on a one-size-fits-all approach, a model must be able to accommodate variation between members of the class idiom, recognizing that different idioms may be processed in different ways depending on their properties. A prototype-based model, where idioms are situated along a continuum of idiomaticity with some phrases more associated with idiomaticity and others less so, may be more reflective of how idioms are perceived and mentally categorized. A prototype-based approach also accounts for the

gradated nature of idiomaticity observed in this work, where idiomaticity is not an all-or-nothing phenomenon but is instead influenced by multiple interacting factors.

Further, these results challenge the traditional reliance on noncompositionality as the defining criterion for idiomaticity. While noncompositionality may play a role in idiom conceptualization, the findings suggest that it is neither necessary nor sufficient to define the construct. Instead, the results suggest a model of idiomaticity in which multiple properties influence perceived idiomaticity**.** Such a model would align with prototype-based theories of categorization, where membership is determined not by strict boundaries but by the degree to which an item exhibits certain prototypical features. This has important theoretical implications. Previous research has often treated idioms as a homogenous class, assuming that they share a set of defining characteristics. However, the presented findings suggest that this is not the case. Instead, idioms vary in their structure and properties. This variation influences how they are perceived and likely extends into comprehension, impacting how they are processed. This points to the need for a more nuanced, prototype-based model of idiomaticity, where idioms are defined not by a single feature but by a combination of properties that together determine their placement on a continuum of idiomaticity.

Moreover, these findings have broader implications for linguistic theory and experimental design. The variability of ratings observed in Experiment 1, and the nuanced role of idiom properties identified in Experiment 2, suggest that idioms occupy a unique position at the intersection of syntax, semantics, and pragmatics. This intersectionality highlights the need for interdisciplinary approaches to the study of idiomaticity, integrating insights from cognitive linguistics, psycholinguistics, and natural language processing. Such approaches are particularly crucial for understanding how idioms are represented and processed in

the mental lexicon and how these processes vary across different contexts and populations. Future research should explore how these findings generalize to other populations, including non-native speakers and individuals with language-related cognitive impairments. Investigating how idiom properties influence recognition and comprehension in such populations would provide valuable insights into shared intuitions and individual differences regarding the construct of idiomaticity. Additionally, longitudinal language acquisition studies documenting the development of a mental category for idioms by considering changes in idiomaticity judgments over time could shed light on the cognitive and experiential factors that shape this construct. These directions would not only deepen our understanding of idioms as linguistic phenomena but also contribute to higher-level discussions about the nature of linguistic categories and the interplay between language, thought, and cognition.

In addition to proposing characteristics of the mental category idiom and providing evidence for the inclusion of three properties in the cognitive definition of these phrases, this work contributes to the understanding of idiom conceptualization. An understanding of idiom conceptualization is vital because without it, the ability to model idiom comprehension is limited.

In this work, *comprehension* has been operationalized as referring to the processes of activating and retrieving the meaning of an idiom, whereas *conceptualization* involves recognizing and mentally categorizing a phrase as an idiom (Langacker 1987). As previously mentioned, comprehension refers to how one engages with an idiom in real-time, involving preconscious processes that allow for understanding without necessarily requiring analysis of individual components. Conversely, conceptualization involves metalinguistic reflection and is concerned with forming, organizing, and structuring knowledge about concepts and mental categories. This

includes idiom analysis as well as categorization based on individual idiom properties and their broader categorization within formulaic language.

Historically, when modeling how idioms are understood, the role of idiom conceptualization has been ignored, validating continued reliance on assumptions about the cognitive category idiom. This line of reasoning is logical and is not limited to idioms as an ability to analyze a word or phrase is not necessarily a requisite for successful comprehension. For example, linguists argue over which semantic role is most appropriate for *lightening* and *the tree* in a sentence such as "*Lightning struck the tree*." Is the agentive aspect of *lightening* more relevant, or does the inanimate aspect of *lightening* result in conceptualization as a cause or force? Is patient, theme, or experiencer more representative of the features most strongly influencing the way one conceives of *the tree*? Crucially, the lack of agreement over the most applicable semantic role does not mean that native speakers struggle to comprehend this sentence, nor do they generally find the role played by these arguments to be ambiguous. This suggests that higher-level analytic understanding is not relied upon during online comprehension, at least when the general syntactic and semantic patterns and combinatoric potentials of lexical items are known.

This line of reasoning extends to idioms. When a phrase is highly familiar, detailed understanding of how its meaning is created is not relied upon to activate the meaning of the phrase. Instead, highly formulaic phrases, irrespective of idiomaticity, are stored on something of a "short cut" list in that they become entrenched. Entrenchment removes the need to recompute meaning each time a phrase is encountered, allowing for quicker meaning retrieval. From this perspective, how one conceives of a phrase (e.g., as idiomatic or non-idiomatic, as having distributed idiomatic meaning or not having distributed idiomatic meaning,

etc.), has little impact on whether a phrase will be successfully comprehended or how comprehension is accomplished (Gibbs 1980). Thus, a deep understanding of idiom conceptualization may not be necessary to model the comprehension of familiar idioms. However, an understanding of idiom conceptualization is necessary to progress beyond this stage. For example, without understanding how idioms are conceptualized, we are unable to explain why native speakers share intuitions about implicit rules for creative idiom use, such as those demonstrated in 1-4 in Table 13. Indeed, current models are unable to account for the often-seamless manner in which novel idioms are comprehended as models would predict failed processing requiring a garden path analysis (Bobrow & Bell 1973, Swinney & Cutler 1979), or slowed processing requiring at least partial reevaluation (cf. Cacciari & Tabossi 1988, Gibbs 1980, 1985, Titone & Connine 1994, Sprenger et al. 2008, Qualls & Harris 2003, Morid et al. 2021). Additionally, the major comprehension models are unable to account for idiom acquisition, instead picking up after one has reached a level of expertise (cf. Bobrow & Bell 1973, Swinney & Cutler 1979, Gibbs 1980, Cacciari & Tabossi 1988, Gibbs et al. 1989, Gibbs 1997, Titone & Connine 1994b, 2014, Sprenger et al. 2006).

Implicit intuitions for allowable usage	
1a. Since the EU referendum, it has been far from clear that people will **always** *pay through the nose* to own property.[28]	1b. You may die a thousand deaths, but you can only **truly** *kick the bucket* once.[29]

Native speakers share intuitions that both 1a. and 1b. can be externally modified.

2a. Leave no **legal** *stone* unturned.[30]	2b. *You may die a thousand deaths, but you can only *kick the **big** bucket once*.[32]
3a. He **may** be *a day late* but he **ain't *no*** *dollar short*.[31]	3b. **By and **very** large*, these therapeutic paradigms focused primarily on adults rather than children.[33]

However, only certain phrases allow for internal modification. Generally, decompositional phrases allow for this type of modification while noncompositional phrases do not. This can be seen in the example above whether the decompositional phrases in 2a. and 3a. allow for modification of a single consituent while the noncompositional phrases in 2b. and 3b. do not.

4a. **Those strings**, he wouldn't *pull* for you.[34]	4b. ***The nose,** people will *pay through*.

Similarly, only certain phrases, usually those that are decompositional, can undergo constituent movement. For example, 4a. allows for fronting of the NP *those strings* while 3b. does not.

[28] savills.de
[29] english-corpus.org
[30] portal.ct.gov
[31] twitter.com
[32] english-corpus.org
[33] dictionary.cambridge.org
[34] Chae 2015, p.50

5a. It was only **a small one** *that the politician spilled*, but **that single bean** ignited a frenzy.	5b. *It was only **a small one** *that he kicked*, but **that single bucket** made us very sad.
Certain phrases are particularly productive. Such phrases tend to be at least decompositional and highly familiar.	

Table 13. Comparison of intuitions regarding allowable instantiations

5.3. Implications

This work has methodological and theoretical implications. The methodological implications are most applicable to the implementation of future work. However, they also have consequences for the interpretation of findings from prior work. The theoretical implications pertain to how idioms are understood and modeled in psycholinguistics and theoretical linguistics. The following subsections discuss these implications.

5.3.1. Methodological Implications

The findings of this research have significant methodological implications for future studies on idiom comprehension. One of the most pressing implications is the need to reconsider how idioms are selected and categorized in experimental work. As demonstrated in Experiment 1, native speakers do not reliably differentiate idioms from non-idiomatic collocations, suggesting the insufficiency of current idiom selection practices based on factors such as dictionary inclusion or personal or small group determinations, even when determinations are made by experts. Instead, future studies must employ more rigorous methods to ensure that the stimuli used to investigate a construct are truly representative of the intended variable.

With respect to the investigation of idiomaticity, one way to create a valid "idiom" condition in which it can be assumed that native speakers generally perceive stimuli as idiomatic is to perform large-scale norming tasks where a representative

sample of native speakers rate the idiomaticity of a number of phrases. Phrases with a high degree of agreement could be retained for study inclusion, ensuring that the phrases used in experiments reflect actual language use, rather than theoretical or prescriptive definitions. Additionally, because properties significantly impact how phrases are perceived, they should be systematically incorporated into experimental designs, either as predictors or as control variables. This gives researchers the option of isolating the effects of idiomaticity or to selectively consider the role of properties so as to better understand the cognitive mechanisms underlying idiom comprehension.

The import of large-scale norming may be easy to minimize. Such an undertaking would be costly and time consuming. Additionally, as experts, a small group of researchers or other area experts seem more than trustworthy. However, this expertise may prove to be problematic as experts are not necessarily representative of the average population. Prior work suggests that the more experience one has with idioms, the more their ratings differ from those with less expertise (Siyanova-Chanturia & Van Lancker Sidtis 2018). Thus, despite the expense, large-scale ratings tasks are necessary prior to any work investigating idiomaticity or properties.

To illustrate this point, a post-hoc analysis was conducted using data collected during piloting and data collected during the course of the real experiment. The analysis assessed the relationship between properties and idiomaticity for experts (pilot phase data) as compared to non-experts (real experiment data). In the piloting phase, 10 linguistics PhD students rated each test phrase as 1.) idiomatic or not idiomatic, 2.) having distributed idiomatic meaning or not having distributed idiomatic meaning, 3.) being partially literal or not being partially literal, and 4.) having a plausible dual literal and idiomatic meaning or not having a plausible dual

literal and idiomatic meaning. When phrases are normed during the experimental design phase, it is common for those aiding in the norming to have some knowledge of the project. In this case, the PhD students had a general idea of the project. Importantly, all had a strong background in linguistic theory regarding idioms and all were familiar with the 3 properties. Thus, unlike naïve, non-expert participants, the PhD students knew the project was about idiomaticity, that each property was thought to be reflective of idiomaticity and approached ratings with a strong personal understanding of the construct of idiomaticity as well as personal strategies for determining whether a phrase exhibited a given property.

Ratings provided by PhD students were calculated, relationships between idiomaticity and properties were assessed, and the distributional frequencies were used to create a continuum of idiomaticity. This continuum was compared with the continuum constructed from the ratings provided by naïve non-expert participants (see Figure 7).

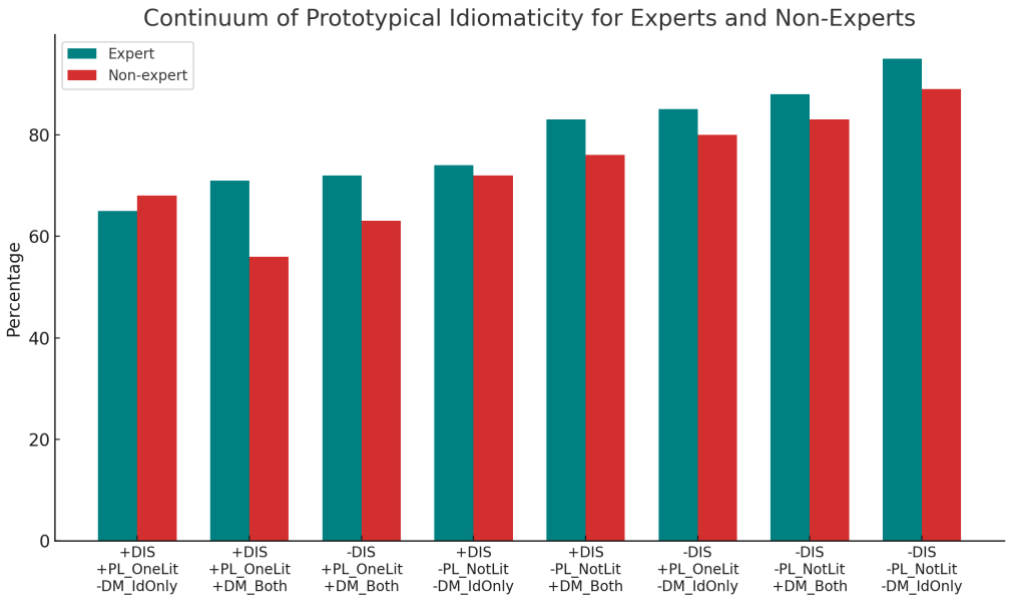


Figure 7. Differences in the continuum of prototypical idiomaticity when constructed based on a small number of expert ratings versus a representative sample of non-expert ratings**.**

In both cases, significant relationships were observed between properties and idiomaticity. However, the strength of these relationships differed. One might wonder if this is due to a lack of statistical power as a sample size of 10 is not sufficiently powered. However, while there is no standard for norming, it is not unusual to collect ratings from 2-3 people in addition to the researcher. Thus, a sample size of 10 is significantly larger than might otherwise be used to inform and refine experimental design.

These differences highlight the importance of careful norming in experimental design. For experiments utilizing categorization tasks, miscategorization of stimuli in the experimental design or failure to control for properties not directly investigated but known to impact idiom conceptualization may invalidate the

results. For example, on the surface, an experiment investigating the impact of distributed idiomatic meaning during online comprehension relies on correct categorization of phrases as 1.) idiomatic versus non-idiomatic and 2.) having distributed idiomatic meaning versus having no distributed idiomatic meaning. If phrases not construed as idiomatic are included in the idiom condition or phrases perceived as having no distributed idiomatic meaning are included in the distributed idiomatic meaning condition, valid conclusions from the experimental findings cannot be drawn.

Further, in light of the experiment 2 finding demonstrating that three idiom properties individually impact idiom conceptualization as well as the finding that certain collections of properties may make a phrase seem more or less idiomatic than may otherwise be expected, failure to control for properties not directly investigated may result in type I and type II errors. Future work should investigate interactions between properties with respect to perceived idiomaticity to better understand the impact of properties so as to gain a more complete understanding of their relationship with idiomaticity. Additionally, future work could carefully reevaluate previously drawn conclusions in light of the potential for contamination by uncontrolled factors. Anecdotal evidence suggests that such reanalysis could be particularly valuable in accounting for the high historic incidence of inconsistent findings, including the common occurrence of failed experiment replication when replication is conducted by a different research group than the original work. Such an analysis has the potential to unify the field, allowing for more valid comparison between experiments as well as the ability to draw on the strengths of different works that may currently be viewed as mutually incompatible.

Finally, there is a need to reconsider how idioms are operationalized in models of language processing. For example, holistic models assume native speakers share

intuitions about members of the class idiom, generally theorizing that idioms are represented and processed as single units in the mental lexicon. The findings challenge this assumption by showing intuitions are not shared for all idioms and that idiom properties influence the degree to which shared intuitions can be assumed. This is in line with individual analysis claims theorizing that idioms are not processed holistically in all cases and emphasizing processing differences for phrases based on the properties they express. Future models should account for the large degree of variation between members of the class idiom as well as the continuous nature of this class.

5.3.2. Theoretical Implications

In addition to methodological considerations, the findings from this research have important theoretical implications for how idioms are understood and modeled in psycholinguistics and linguistics. Due to knowledge gaps surrounding idiom conceptualization, a cognitive definition of idiomaticity has not been established. Instead, a working definition has been adopted from linguistics with little concern for the originally intended application of the adopted definition. For example, a generative grammarian might adopt a straightforward, dichotomous definition of idioms as non-compositional phrases to simplify the modeling of syntactic theories. This approach makes sense when the goal is to create syntactic sketches or language descriptions. as defining idioms as fixed, noncompositional units simplifies the required account of idioms, allowing for the treatment of idioms as stable entries stored in the lexicon, separate from productive syntactic processes. While such an approach may be useful when focusing on formal descriptions of language as it prioritizes structural regularity, the approach does not take cognitive mechanisms into account. Consequently, adoption of a strictly dichotomous view may limit insights into idiom comprehension, where degrees of compositionality

may play a role in how idioms are recognized, processed, or interpreted. Despite this risk, holistic approaches have largely adopted a dichotomous view from the generative tradition, without fully considering the implications of repurposing this definition for psycholinguistic work and without validation to be sure this adopted definition adequately defines the class of phrases intended, and believed, to be under investigation or that such a definition is appropriate to model cognition (cf. Bobrow & Bell 1973, Swinney & Cutler 1979). By adopting a dichotomous definition suited to syntactic or lexical descriptions, holistic models implicitly assumed the definition would also capture how idioms are mentally processed.

Knowledge of which properties impact comprehension should inform theories of idiom conceptualization – properties of idioms that also impact comprehension should be evaluated as possible defining features of the class idiom that may also serve as indicators of prototypicality. Similarly, an established description of the class idiom should inform theories of comprehension, featuring prominently in experimental design and driving predictions as properties that define the class idiom and moderate a phrase's degree of prototypicality are likely to impact comprehension. Thus, this work addressed the lack of a shared, more complete description of the class idiom by first testing, then rejecting, a dichotomous definition and second by taking the first steps towards a systematic evaluation of properties in order to establish a tested description of the cognitive class idiom.

5.4. Limitations and future directions

As with any research, this study is not without its limitations, and these limitations provide important directions for future research. One of the primary limitations lies in the use of metalinguistic judgments to assess idiomaticity. While metalinguistic tasks are valuable for probing explicit awareness and analysis of language, they are not necessarily reflective of how idioms are processed in real-time during everyday

language use. Metalinguistic judgments capture how participants consciously analyze and categorize phrases rather than how they process language automatically in natural contexts. In this sense, such tasks may reflect "how one thinks they understand" language rather than "how one actually understands" language.

This distinction has significant implications. If analytic, reflective tasks involve different processes than those relied upon during natural language use, then the results of metalinguistic tasks may not fully capture the cognitive mechanisms involved in idiom comprehension. In contrast, online tasks that measure reaction times and other indirect indicators of processing may offer a more accurate representation of the automatic, subconscious processes involved in idiom recognition. However, such tasks rely on proxy measures of idiomaticity, assuming measures such as time are sufficiently reflective of idiomaticity so as to neutralize the impact of inherent random effects. Future work should complement metalinguistic tasks with online measures, such as reaction time, eye-tracking, or EEG studies to gain a more complete understanding of idiom processing. Combining these methodologies allows for a more comprehensive assessment of the relationship between idiom properties and the cognitive processes underlying comprehension, thus bridging the gap between "how one thinks they understand" and "how one actually processes" idiomatic language.

A second limitation of this study is the focus on a finite set of idiom properties. The goal of this research was to lay the groundwork for a complete description of the cognitive category idiom. While this research examined the role of formulaicity, distributed idiomatic meaning, partial literality, and dual meaning in idiom perception, there are other properties that are likely to impact perceived idiomaticity, such as flexibility, transparency, and figuration. Future work should

build on these findings by considering the role of these and other idiom properties in idiom understanding as well as considering whether there are interactions between idiom properties that impact their relationship with perceived idiomaticity.

Another potential limitation to generalizability lies in the structure of the idioms and non-idiomatic collocations included in this work, which was limited to two-word noun and adjective-headed phrases. While this was necessary to control for factors such as phrase length and formulaicity, it limits the generalizability of the findings. Future research should consider a wider variety of idiom types, including longer phrases and verb-headed idioms, to determine whether the patterns observed in this study hold across different types of idioms.

Additionally, future work should more carefully consider the relationship between dual meaning and transparency. In this work, dual meaning was operationalized in a manner that excluded a consideration of the role of transparency. While these constructs should not be conflated, a strong relationship between them is suspected. Phrases with no dual meaning were more predictive of idiomaticity than those with dual meaning. However, dual meaning should be used with caution due to a complex relationship with transparency. When a phrase has an opaque dual meaning, or a meaning that does not have an obvious relationship with a literal interpretation, the phrase stands out more, marking it as more unusual and making it easier to recognize as an idiom. However, when the non-literal meaning of a phrase is transparently related to its literal meaning or when its nonliteral meaning is a small extension of the literal meaning, the phrase may stand out less, making it more difficult to recognize as an idiom. Future work should consider whether idiom prototypicality differs for phrases with transparently versus opaquely related dual meaning.

A final limitation of this work lies in its reliance on a more traditional prototype categories approach. Future work could carefully consider a hybrid approach to idiomaticity that incorporates aspects of a prototype approach and aspects of a radial categories prototype approach (radial approach). There are several reasons that may make a more traditional prototype approach seem less than ideal. One might wonder why a continuous, fuzzy set theory, or radial category prototype approach was not adopted. Continuous and fuzzy set approaches allow for partial category membership, where a phrase may be "somewhat idiomatic" while prototype and radial approaches assume full membership but with graded typicality, where all members are idioms but may be stronger or weaker examples of an idiom (Zadeh 1965, Lakoff 1972, 1987, 2007, Rosch 1978, Langacker 1987, 2000, 2014, Croft 2001, Evans et al. 2007, Taylor 2009, Bybee 2010). Given the findings of experiment 1, which demonstrated the existence of a category idiom despite variation in the strength to which individual members were recognized as strong representatives, continuous and fuzzy set approaches are not an ideal fit for this data.

Within the prototype framework, a somewhat alternative approach would be to take a radial stance, which is viewed as a refined version of a prototype approach. On the surface, radial categories seem to provide a method ideally suited to account for the continuum of perceived idiomaticity. A radial approach allows for fuzzy boundaries, accounts for motivated extensions, and models degrees of similarity to a central prototype, all of which the prototype approach has been criticized for lacking explicit mechanisms to formalize. However, despite these apparent strengths, a closer look reveals fundamental mismatches between radial categories and the data presented here. Moreover, the data fails to meet preconditions of a radial approach, such as the existence of a central subcategory member or the fact that category membership is not strongly predictable.

One seeming advantage of a radial category approach is its more sophisticated category structure, including an ability to model multifaceted relationships and provide a more structured account of category boundaries. Perhaps the biggest difference between a prototype and a radial approach is that a radial approach moves beyond the simple scalar membership system seen with a prototype approach, offering explanations for phenomena such as fuzziness and typicality effects.

In a prototype approach, categories are organized around a central exemplar, with membership based on degrees of similarity. As such, membership is gradated, based on similarity to a central member. This has led some to criticize the prototype approach, suggesting it oversimplifies category structure while simultaneously obfuscating category boundaries. However, Lakoff (2007) notes that oversimplification is a misconception which assumes categories are represented in the mind solely in terms of prototypes and results from a narrow interpretation of prototype theory that was never intended (Rosch 1987, Lakoff 1987). Focusing on the degree to which an entity is representative of a given category can overshadow additional factors that enrich a cognitive model. Instead, Lakoff (2007) argues that prototype categories can have internal structure beyond simple resemblance, incorporating structured relationships that go beyond mere similarity to a central exemplar.

Taking this a step further, Lakoff notes that misconceptions extend to an assumption that goodness-of-example ratings are scalar whenever a category is scalar (Lakoff 1987). This assumption adopts a view of categories as being represented in the mind solely in terms of prototypes and implies that all category membership is graded, rather than recognizing that membership can be binary while typicality is graded. This can be seen in a common interpretation of

Armstrong et al. (1983), who found that native speakers rated common, short numbers, such as two and three, as "more representative" of the categories "even" and "odd" than other, longer numbers, such as 806 or 447. Logically, these categories are not gradient; membership is binary. If this work is interpreted through the lens of gradient membership, no prototype effects should be seen for evenness or oddness. Thus, under this interpretation in which prototypes constitute mental categories, the work of Armstrong et al. (1983) would demonstrate an assumption fallacy underlying the prototype account, the effects as structural interpretation fallacy, which holds that effects characterize the structure of a category as it is represented in the mind (Lakoff 1987:43). According to this assumption, goodness effects in judgments should be seen if and only if category membership is scalar. Crucially, this is not the interpretation of prototype theory intended by Rosch (1978) or Lakoff (1987, 2007), as Rosch (1978) explicitly states that effects constrain representations but do not directly correspond to them in a one-to-one manner.

In line with the interpretation of prototype theory intended by Rosch and Lakoff (1987), an alternative explanation of Armstrong et al. (1983) holds that their work uncovered an inherent tendency to incorporate basicness, frequency, salience, commonality, and other such features into the perception of category members. The fact that prototype effects emerge in a fully discrete category demonstrates that graded typicality can exist independently of gradient membership. Crucially, perception does not change the structure of an item. Prototypes are not theories of representation for categories, nor should they be viewed as reflective of processing in any way. "To speak of a prototype at all is simply a convenient grammatical fiction; what is really referred to are judgments of degree of prototypicality." (Lakoff 1987:44). Two and 222 are equally even. However, one would expect a difference in

the degree of immediate reaction to evenness, the degree to which one incorporates evenness into future thought patterns about the number (i.e., activates the concept of evenness), and the degree to which one conceives of a number as a member of the class "even" based on prior experience with the specific number. Rather than exposing a flaw in prototype theory, these findings illustrate how prototype-driven effects can emerge even in non-scalar categories, reflecting cognitive biases in how category members are perceived rather than how they are defined.

While prototypes account for graded typicality effects, they do not inherently distinguish between structural category membership and cognitive salience. This gap contributed to skepticism about prototype models but also underscored the need for a more nuanced framework, one that could accommodate both categorical structure and variation in cognitive accessibility. Radial category models address this by modeling structured variation within a category based on relationships to the central prototype. In this way, a radial approach can help to clarify how prototypicality effects emerge in non-scalar categories like "even" and "odd".

In light of this, a radial approach to the data presented in this dissertation seems advantageous. While a prototype approach accommodates gradient membership, it cannot account for the complex structure motivating membership beyond prototypicality (Rosch 1978). By using radial categories, it would seem as though the structure of relationships between subcategories could be included in the model as multiple dimensions could be considered and explicitly factored into the internal model structure. However, while not incompatible, radial categories are less equipped to account for gradient membership and are ill-equipped to handle many common types of relationships. Specifically, any category in which nonlinear relationships dictate boundaries, membership, or distance from the central subcategory, making a radial approach a less-than-ideal choice for this data.

Additionally, a radial approach requires a central subcategory member, defined as the case in which the otherwise separable properties that define that member converge. It is unclear from this data that a central subcategory member exists, meaning that it does not meet this precondition.

While a prototype approach may be a more appropriate choice than a purely radial approach with respect to category representation, this is not the only area in which a radial approach offers potential advantages over a prototype approach. In addition to within-category structure, radial categories also improve upon a prototype approach with respect to their ability to account for category boundaries. While membership based on the single dimension of prototypicality oversimplifies, prototype categories often exhibit fuzzy boundaries, making it difficult to determine definitive membership. Although some members are clearly within the category, others are more ambiguous. This fuzziness contrasts with the classical view of categories which assumes sharply delineated boundaries. While fuzziness can be a strength, the prototype approach has been criticized for an inability to provide category boundaries, particularly as category membership may be context dependent and therefore unstable. It also struggles to model categories that have necessary and sufficient conditions or multiple dimensions of structure, which requires an ability to model the impact of multiple features. This is problematic given the multidimensional nature of idiomaticity. One solution is to create a cluster model in which each category represents a feature bundle (e.g., [+DIS, +PL, +DM] or [+DIS, -PL, -DM], situating the categories along a single, feature-based axis. However, radial categories offer a distinct advantage in that they are able to represent multidimensional influences.

In a radial approach, there is no single cognitive model that represents an entire category. Instead, subcategories radiate outward from a central prototype, forming

structured, non-arbitrary links that are motivated by linguistic, cognitive, and cultural factors (Lakoff, 1987; Langacker, 1988). These connections are not purely based on degrees of resemblance but rather on conventionalized extensions. While radial categories maintain a structure based on prototypicality, in that they recognize that members have varying degrees of similarity to the prototype, they impose additional organization through motivated, hierarchical relationships (Lakoff 1987, Langacker 1988). These relationships are conventionalized and are not predictable (Lakoff 1987), just as the pattern of polysemy associated with a word lemma is not predictable. It is important to note that the boundaries of any given radial category are linguistic: a *fruit salad* is considered to be a kind of salad by American English speakers because the noun *salad* happens to be used to refer to it.

Because of these patterns of structured extension, radial categories might seem like an ideal way to model idiomaticity. However, radial categories are poorly equipped to handle gradience when there is no central subcategory, which is a core issue in perceived idiomaticity. Unlike prototype categories, which can accommodate effects of relative goodness-of-example, radial categories rely on discrete subcategories linked by convention rather than scalar membership. This structure is insufficient to capture the continuum of perceived idiomaticity, as it assumes that idioms belong to structured, motivated subcategories rather than forming a flexible, gradable continuum.

Finally, a prototype approach depends on intuitions and has been criticized for being metalinguistic. The fact that this potential issue is addressed by a radial approach by grounding category extensions in non-arbitrary links to provide a more principled account of categorization than one based solely on intuition has proven attractive (Taylor 2004). However, given that this work is inherently based on

intuitions, the metalinguistic underpinnings of a prototype approach are not a weakness but a reasonable and appropriate aspect of the framework in this context.

Overall, the findings of this research suggest that relationships between perceived idiomaticity and idiom properties are structured and predictable, making a prototype-based approach more appropriate than a radial category model. Prototype theory accounts for graded typicality by modeling category membership as a continuum based on feature overlap with a central prototype. The results of this study indicate that idiomaticity judgments are systematically influenced by linguistic properties such as distributed idiomatic meaning, partial literality, and dual meaning, suggesting that perceived idiomaticity is not arbitrary but emerges from structured variation. Unlike prototype categories, radial categories assume that membership extensions are driven by systematic conceptual shifts, such as metaphorical or functional relationships. This assumption is not supported by the findings, as perceived idiomaticity does not appear to be structured by historical or functional extensions but rather by intrinsic linguistic features. Moreover, while prototype theory does not impose strict categorical boundaries, it allows for probabilistic predictions of category membership based on graded feature similarity. The predictability observed in the present findings suggests that idiomaticity is not merely a conventionally assigned label but a category with systematic internal structure. Because the results indicate that perceived idiomaticity is determined by measurable linguistic properties rather than conceptual extensions, prototype theory provides a more suitable framework than a radial model for capturing the structure of idiomaticity as reflected in speaker judgments.

However, neither existing framework is ideally suited to account for the relationship between the investigated idiom properties and perceived idiomaticity. Perceptions arise from native speaker intuitions and require no explicit knowledge,

formal learning, or instruction on how to deal with edge cases. They arise naturally from experience and, as such, may differ slightly between speakers, just as category boundaries for "furniture" may differ between individuals (e.g., whether a shelf is considered furniture, and if so, whether a wall-mounted shelf or a framed picture would also qualify). The fact that properties of idiomaticity are fully predictable should make them easier to model. However, neither a radial nor a prototype approach is fully able to account for this. Predictable membership precludes idioms from a radial approach and, while a prototype approach is not incompatible, it is unable to account for membership. This may slightly advantage a prototype approach. However, while the adopted clustering method groups multidimensional features in a way that allows the category to be treated as if it follows a single continuum, a prototype approach has a further disadvantage in that it is unable to model multiple dimensions. Thus, modeling predictable, multidimensional category membership poses a challenge for the practical implementation of radial and prototype category models.

A further challenge lies in independence. A pre-condition of prototype models is that features contribute independently to category structure, whereas a core assumption of radial models is that features may be interdependent. The properties in this work were statistically independent, which aligns with the expectations of a prototype model. However, idiomaticity judgements were best described when considering combinations of features, suggesting that category membership may rely on the way in which multiple properties jointly impact perception. This complicates a strict interpretation of feature independence and highlights the importance of considering whether moderating relationships between features play a role in perceived idiomaticity. Future work could consider a hybrid approach that incorporates aspects of both approaches, while also building upon them so as to account for

predictable membership within a gradient structure, relationships between categories that may be non-linear, and fuzzy category boundaries.

Despite these limitations, this dissertation provides a robust foundation for understanding the nuanced relationship between idiomaticity and its defining properties, challenging existing assumptions about idiom processing and conceptualization. By bridging theoretical and empirical perspectives, this work not only advances our understanding of idiomatic language but also lays the groundwork for future investigations into considering broader cognitive and clinical implications.

REFERENCES

1. Armstrong, S. L., Gleitman, L., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13,* 263–308

2. Athanasopoulos P., Bylund, E., Montero-Melis, G., Damjanovic, L., Schartner, A., Kibbe, A., Riches, N., Thierry, G. (2015). Two languages, two minds: Flexible cognitive processing driven by language of operation. *Psychological Science, 26*(4), 518-526.

3. Barkema, H. (1996). Idiomaticity and terminology: A multi-dimensional descriptive model. *Studia Linguistica, 50*, 125–160.

4. Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods, 42*(3), 665-670.

5. Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology, 7,* 1116-1127.

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

7. Booij, G. E. (2013). Morphology in Construction Grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0014

8. Bridges, K. A., & Van Lancker Sidtis, D. (2013). Formulaic language in Alzheimer's disease. *Aphasiology, 27*(7), 799-810.

9. Bulkes, N. Z., & Tanner, D. (2017). "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods, 49*, 772-783.

10. Bybee, J. (2010). *Language, Usage and Cognition.* Cambridge University Press.

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

12. Cacciari, C., Corrardini, P., & Ferlazzo, F. (2018). Cognitive and personality components underlying spoken idiom comprehension in context. An exploratory study. *Frontiers in Psychology*, *9*, 659.

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

14. Cain, K., Oakhill, J., & Lemmon, K. (2005). The relation between children's reading comprehension level and their comprehension of idioms. *Journal of Experimental Child Psychology, 90*(1), 65-87.

15. Canal, P., & Bambini, V. (2012). Pragmatics electrified. In *Language Electrified: Principles, Methods, and Future Perspectives of Investigation.* 583-612. New York, NY: Springer US.

16. Carrol, G. (2015). *Found in translation: a psycholinguistic investigation of idiom processing in native and non-native speakers.* Doctoral dissertation. University of Nottingham.

17. Carrol, G., & Conklin, K. (2019). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech, 63*(1), 95-122.

18. Carrol, G. (2023). Old Dogs and New Tricks: Assessing Idiom Knowledge Amongst Native Speakers of Different Ages. *Journal of Psycholinguistic Research*, *52*(6), 2287-2302.

19. Casasanto, D. & Lupyan, G. 2015). All Concepts are Ad Hoc Concepts. In *The Conceptual Mind: New directions in the study of concepts*. E. Margolis & S. Laurence (Eds.), 543-566. Cambridge: MIT Press.

20. Cavanaugh, J., & Neath, A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley interdisciplinary reviews. *Computational Statistics, 11*(3). DOI: 10.1002/wics.1460.

21. Chafe, W. L. (1970). *Meaning and the structure of language*. University of Chicago Press.

22. Chakrabarty, M., Klooster, N., Biswas, A., & Chatterjee, A. (2023). The scope of using pragmatic language tests for early detection of dementia: A systematic review of investigations using figurative language. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association.*

23. Church, K.W. and Hanks, P. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics, 16*(1), 22-29.

24. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

25. Company, Concepción. (2012). Historical Morphosyntax and Grammaticalization. In *The Handbook of Hispanic Linguistics ed*. Hualde, J., Olarrea, A., & O'Rourke, E. 673–693. London/New York: Blackwell. 10.1002/9781118228098.ch31.

26. Coulson, S., & Van Petten, C. (2002). Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition, 30*(6), 958-968.

27. Croft, W. (2001) *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

28. Croft, W., & Cruse, A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

29. Cronk, B. C., & Schweigert, W. A. (1993). The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics, 13*(2), 131-146.f

30. DATAtab: DataTab Team (2024). DataTab: Online Statistics Calculator. DataTab e.U. Graz, Austria. URL https://datatab.net.

31. Davies, M. (n.d.). *Association Measures.* https://www.english-corpora.org/help/association-measures.pdf.

32. Davies, M. (2008). *The Corpus of Contemporary American English* (COCA). Available online at https://www.english-corpora.org/coca/.

33. Davies, M. (2008-). *Collocates data from The Corpus of Contemporary American English* (COCA).

34. Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory, (6)*2.

35. de Silva, A. (2018). *Context Effects on Ambiguous Idiom Comprehension in Older and Younger Adults.* University of Maine.

36. Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *(5)*1, 61-78. https://doi.org/10.1515/CLLT.2009.003

37. Evans, V. (2007). *Glossary of cognitive linguistics.* Edinburgh University Press.

38. Evert, S. (2008). Corpora and collocations. Extended manuscript https://lexically.net/downloads/corpus_linguistics/Evert2008.pdf. Adapted from A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, *58*. Mouton de Gruyter, Berlin.

39. Fanari, R., Cacciari, C., & Tabossi, P. (2010). The role of idiom length and context in spoken idiom comprehension. *European Journal of Cognitive Psychology, 22*(3), 321-334.

40. Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press

41. Fazly, A., & Stevenson, S. (2006). Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations. In *11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 337-344).

42. Fernando, C. (1996). *Idioms and Idiomaticity*. Oxford: Oxford University Press.

43. Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage publications limited. ISBN-13 978-9351500827.

44. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

45. Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language, 6,* 22-42.

46. Gazzaniga, M. S. (Ed.). (2014). *Handbook of Cognitive Neuroscience*. Springer.

47. Geeraert, K., Newman, J. and Baayen, R.H. (2017). Idiom Variation: Experimental Data and a Blueprint of a Computational Model. *Topics in Cognitive Science, 9,* 653-669. DOI: 10.1111/tops.12263.

48. Geeraerts, D. (1989). Introduction: Prospects and Problems of Prototype Theory. *Linguistics*, *27*(4), 587-612. https://doi.org/10.1515/ling.1989.27.4.587.

49. Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology, 113*, 256-281.

50. Gibbs, R. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition, 8,* 149-156.

51. Gibbs, R., Nayak, N. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21*, 100-138.

52. Gibbs, R. W., Jr., Nayak, N. P., Bolton, J. L., & Keppel, M. E. (1989). Speakers' assumptions about the lexical flexibility of idioms. *Memory & Cognition, 17*, 58-68.

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

54. Gibbs, R. W., & O'Brien, J. E. (1990). Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. *Cognition, 36*(1), 35-68. DOI: 10.1016/0010-0277(90)90053-M.

55. Gibbs, R. (1990). Psycholinguistic studies on the conceptual basis of idiomaticity. *Cognitive Linguistics, 1*, 417-451.

56. Gibbs, R. (1991). Semantic analyzability in children's understanding of idioms. *Journal of Speech and Hearing Research, 34*, 613-620.

57. Gibbs, R. W. (1993). Why idioms are not dead metaphors. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure and interpretation,* 57–78. Hillsdale, NJ: Erlbaum.

58. Gibbs, R. (1994). *The Poetics of Mind*. Cambridge University Press, Cambridge.

59. Gibbs, R. W., Bogdanovich, J. M., Sykes, J. R., & Barr, D. J. (1997). Metaphor in idiom comprehension. *Journal of Memory and Language, 37*(2), 141-154. DOI: 10.1006/jmla.1996.2506.

60. Gibbs, R. W. (2011). Evaluating Conceptual Metaphor Theory. *Discourse Processes, 48*(8), 529-562. DOI: 10.1080/0163853X.2011.606103.

61. Gibbs, R. W. (2020). Personal communication. August 2020.

62. Glucksberg, S., Brown, M., & McGlone, M. S. (1993). Conceptual metaphors are not automatically accessed during idiom comprehension. *Memory & Cognition, 21*(5), 711–719. https://doi.org/10.3758/BF03197201.

63. Glucksberg, S. (2001). *Understanding Figurative Language: From Metaphors to Idioms*. Oxford University Press, Oxford.

64. Hockett, C. F. (1958). *A Course in Modern Linguistics*. New York: Macmillan.

65. Hunston, S. (2002). Pattern grammar, language teaching, and linguistic variation. *Using Corpora to Explore Linguistic Variation,* 167-183.

66. IBM Corp. (2020). *IBM SPSS Statistics for Windows, Version 29.0.* (Computer software). IBM Corp.

67. Jesussek, M. & Volk-Jesussek, H. (2024). Statistics made easy. 4th edition. DATAtab e.U. Graz. 2024.

68. Jurafsky, D., and James H. M., (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. (Online manuscript released January 12, 2025.) https://web.stanford.edu/~jurafsky/slp3/.

69. Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129-191.

70. Kay, P., & Michaelis, L. A. (2012). Constructional meaning and compositionality. *Semantics: An international handbook of natural language meaning*, *3*, 2271-2296.

71. Kędzierska, H. (2018). *The Role of Context in the Processing of Idioms: An Experimental Study*. Center for General and Comparative Linguistics.

72. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*, 7-36.

73. Kilgarriff, A., & Kosem, I. (2012). *Corpus Tools for Lexicographers*. 31-55. na.

74. 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.

75. Kučera, H., Francis, W., Twaddell, W. F., Marckworth, M. L., Bell, L. M., & Carroll, J. B. (1967). *Computational Analysis of Present-Day American English*. RI: Brown University Press.

76. Lakoff, G. (1972). Hedges: A study in meaning criteria and the logic of fuzzy concepts. In *Papers from the Eighth Regional Meeting of the Chicago Linguistics Society*. Also in *Journal of Philosophical Logic* (1973), 2, 458–508.

77. Lakoff, G. (1987). Women, fire, and dangerous things. University of Chicago Press, Chicago.

78. Lakoff, G. (2007). Cognitive models and prototype theory. In V. Evans, B. Bergen, & J. Zinken (eds.), *The cognitive linguistics reader,* 130-167. Equinox Publishing Ltd.

79. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

80. Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites*. Stanford university press.

81. Langacker, R. W. (2000). *Grammar and conceptualization*. Walter de Gruyter.

82. Langacker, R. W. (2013). *Essentials of cognitive grammar*. Oxford University Press, Incorporated.

83. Langacker, R. W. (2014). Conceptualization, symbolization, and grammar. In *The new psychology of language,* 1-37. Psychology Press.

84. Levorato, M. C., and Cacciari, C. (1995). The effects of different tasks on the comprehension and production of idioms in children. *Journal of Experimental Child Psychology, 60*, 261–283. DOI: 10.1006/jecp.1995.1041.

85. Levorato, M. C., and Cacciari, C. (2002). The creation of new figurative expressions: psycholinguistic evidence in Italian children, adolescents and adults. *Journal of Child Language, 29*, 127–150. DOI: 10.1017/s0305000901004950.

86. Libben, M., Titone, D. (2008). The multidetermined nature of idiom processing. *Memory & Cognition, 36,* 1103-21. DOI: 10.3758/MC.36.6.1103.

87. Libben, M., & Titone, D. (2011). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon*, *9*(3), 473-496.

88. Lindholm, C., & Wray, A. (2011). Proverbs and Formulaic Sequences in the Language of Elderly People with Dementia. *Dementia*, *10*(4), 603-623.

89. Longman ID (1998). *Longman idiom dictionary*. Harlow: Addison Wesley Longman Limited.

90. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

91. McGlone, M. S., Glucksberg, S., & Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse processes*, *17*(2), 167-190.

92. McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276–282.

93. Mel'cuk, I. (1995). Phrasemes in language and phraseology in linguistics. In Everaert et al. (Eds.), *Idioms,* (167-232). Hillsdale N.J.: Lawrence Erlbaum Associates.

94. Michaelis, L. (2013). Sign-Based Construction Grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

95. Michaelis, L. A. (2017). Meanings of Constructions. *Oxford Research Encyclopedia of Linguistics.* DOI: 9780199384655.013309.

96. Michaelis, L. A. (2019). Constructions are Patterns and So Are Fixed Expressions. In B. Busse & R. Moehlig-Falke (Eds.), *Patterns in Language and Linguistics: New Perspectives on a Ubiquitous Concept*. Berlin, Boston: De Gruyter Mouton. 193-220. https://doi.org/10.1515/9783110596656-008

97. Mueller, R. A., & Gibbs, R. W. (1987). Processing idioms with multiple meanings. *Journal of psycholinguistic research*, *16*, 63-81.

98. Nenonen, M. (2007). Prototypical idioms: evidence from Finnish. *SKY Journal of Linguistics*, *20*, 309–330.

99. Nippold, M. A., & Taylor, C. L. (2002). Judgments of idiom familiarity and transparency. *ASHA.*

100. Nordmann, E., Cleland, A. A., & Bull, R. (2014). Familiarity breeds dissent: Reliability analyses for British-English idioms on measures of familiarity, meaning, literality, and decomposability. *Acta psychologica*, *149*, 87-95.

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

102. Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of verbal learning and verbal behavior*, *17*(4), 465-477.

103. Papagno, C. (2011). Idiomatic language comprehension: Neuropsychological evidence. *Neuropsychology of communication,* 111-129. Milano: Springer Milan.

104. Penttila, E. (2010). Prototype-based taxonomy of idiomatic expressions. *Applications of Cognitive Linguistics*, *14*, 145–162.

105. Pitt, D., & Katz, J. J. (2000). Compositional Idioms. *Language, 76*(2), 409-432. https://doi.org/10.2307/417662.

106. Qualls, C. D., Obler, L. K., Connor, L. T., & Albert, M. L. (2001, April). Idiom and proverb interpretation: An indication of metaphoric processing in

aging. Paper presented at the *Fourth Annual Conference on Researching and Applying Metaphor (RAAM IV)*, Tunis, Tunisia (North Africa).

107. Qualls, C. D., & Harris, J. L. (2003). Age, working memory, figurative language type, and reading ability. *ASHA Research*.

108. Qualls, C. D., O'Brien, R. M., Blood, G. W., & Hammer, C. S. (2003). Contextual variation, familiarity, academic literacy, and rural adolescents' idiom knowledge. *Language, Speech, and Hearing Services in Schools, 34*(1), 69–79.

109. Rapp, A. M., Mutschler, D. E., & Erb, M. (2012). Where in the brain is nonliteral language? A coordinate-based meta-analysis of functional magnetic resonance imaging studies. *Neuroimage, 63*(1), 600–610.

110. Ramberg, C., Ehlers, S., Nydén, A., Johansson, M., & Gillberg, C. (1996). Language and pragmatic functions in school-age children on the autism spectrum. *International Journal of Language & Communication Disorders, 31*(4), 387-413.

111. Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, *25*(5), 762-776.

112. Rosch, E. H. (1973). Natural categories. *Cognitive psychology, 4*(3), 328-350.

113. Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology, 7*(4), 573-605.

114. Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (eds), *Cognition and categorization.* Hillsdale, NJ: Erlbaum

115. Rosch, E., & Lloyd, B. B. (Eds.). (2024). *Cognition and categorization*. Taylor & Francis.

116. Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In *RASLAN*, 6-9.

117. Sanchez-Lopez, E., (2015). Phraseologization as a process of semantic change. *Catalan Journal of Linguistics, 14*, 159-177.

118. Schweigert, W. A. (1985). The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research, 15,* 33-45.

119. Schweigert, W. A., & Moates, D. R. (1988). Familiar idiom comprehension. *Journal of psycholinguistic research, 17,* 281-296.

120. Schweigert, W. A., & Cronk, B. C. (1992). Ratings of the familiarity of idioms' figurative meanings and the likelihood of literal meanings among US college students. *Current Psychology, 11,* 325-345.

121. Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 2*, 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

122. Shaoul, C., Baayen, R. H., & Westbury, C. F. (2014). N-gram probability effects in a cloze task. *The mental lexicon, 9*(3), 437-472.

123. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological bulletin*, *86*(2), 420.

124. Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*(2), 251-272.

125. Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics, 36*, 549–569

126. Sprenger, S., Levelt, W., Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journa*l of Memory and Language, 54, 161–184.

127. Sprenger, S. A., la Roi, A., & Van Rij, J. (2019). The development of idiom knowledge across the lifespan. *Frontiers in Communication, 4*(29).

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

129. Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London etc.: Sage Publishers. ISBN 9781849202008.

130. Tabossi P, Fanari R, Wolf K. (2008). Processing idiomatic expressions: effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(2), 313-27. DOI: 10.1037/0278-7393.34.2.313. PMID: 18315408.

131. Tabossi, P., Fanari, R. & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition, 37*, 529–540. https://doi.org/10.3758/MC.37.4.529.

132. Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

133. Taylor, J. R. (2009). *Linguistic categorization.* Oxford Univ. Press. ISBN: 978-0-19-926664-7. OCLC: 553516096.

134. Thibodeau, P. H., Sikos, L., & Durgin, F. H. (2018). Are subjective ratings of metaphors a red herring? The big two dimensions of metaphoric sentences. *Behavior Research Methods, 50,* 759-772.

135. Thyab, R. A. (2016). The Necessity of idiomatic expressions to English Language learners. *International Journal of English and Literature*, *7*(7), 106-111.

136. Titone, D. A., & Connine, C. M. (1994a). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor & Symbolic Activity, 9*, 247-270.

137. Titone, D. A., & Connine, C. M. (1994b). Comprehension of idiomatic expressions: Effects of predictability and literality. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 1126-1138.

138. Titone, D. & Connine, C. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics, 31*(3), 1655-1674.

139. 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.

140. 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.

141. Van de Cruys, T. (2011). Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality,* (16-20).

142. Vega Moreno, R. E. (2007). *Creativity and Convention: the pragmatics of everyday figurative speech*. John Benjamins B.V.

143. Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience, 22*(8), 1682-1700.

144. Vilkaite, L. (2016). Are nonadjacent collocations processed faster? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1632–1642. doi: 10.1037/xlm0000259.

145. Vulchanova, M., Vulchanov, V., & Stankova, M. (2011). Idiom comprehension in the first language: a developmental study. *VIAL, 8*, 141–163.

146. Vulchanova, M., Saldaña, D., Chahboun, S., & Vulchanov, V. (2015). Figurative language processing in atypical populations: the ASD perspective. *Frontiers in human neuroscience, 9*(24). https://doi.org/10.3389/fnhum.2015.00024.

147. Vulchanova, M., Milburn, E., Vulchanov, V., & Baggio, G. (2019). Boon or burden? The role of compositional meaning in figurative language processing and acquisition. *Journal of Logic, Language and Information, 28,* 359-387.

148. Wasow, T., Sag, I., Nunberg, G. (1984). Idioms: An interim report. In Hattori, S. Inoue, K. (Eds.), *Proceedings of the XIIIth International Congress of Linguistics.* Tokyo.

149. Weinreich, U. (1969). Problems in the analysis of idioms. *Substance and structure of language,* 23-82. University of California Press.

150. Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

151. Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1-28.

152. Wulff, S. (2008). *Rethinking Idiomaticity*. London: Bloomsbury Publishing. ISBN: 9781441184955.

153. Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8,* 338–353

154. Zheng, H., Bowles, M. A., & Packard, J. L. (2022). NS and NNS processing of idioms and nonidiom formulaic sequences: What can reaction times and think-alouds tell us? *Applied Psycholinguistics, 43*(2), 363-388.

155. B Zimmerer, V. C., Wibrow, M., & Varley, R. A. (2016). Formulaic language in people with probable Alzheimer's disease: A frequency-based approach. *Journal of Alzheimer's Disease, 53*(3), 1145-1160.

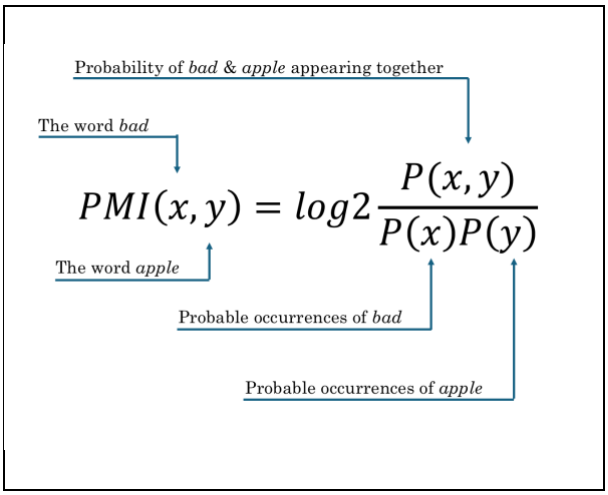
APPENDIX I

CALCULATION OF PMI

This appendix provides additional information about PMI and its implementation in this dissertation, including the equation used for calculating PMI values, factors motivating the use of this metric over other options, a description of the calculation process, and an evaluation of PMI values returned by the PMI script. It concludes with a discussion of limitations and considerations should this script be used to calculate PMI for other types of projects

1. Pointwise mutual information

As previously stated, pointwise mutual information (PMI) is an information theoretic metric that estimates the degree of phrasal formulaicity by measuring the likelihood that two words will appear together intentionally rather than by chance (Fano 1961, Church & Hanks 1990, Manning & Schuetze 2000). It is commonly used in natural language processing, with applications for tasks such as automatic idiom recognition (cf. Lin 1999, Fazly & Stevenson 2006). It has also been used in psycholinguistics to quantify the degree of formulaicity for idioms and non-idiomatic collocations (e.g., *ice cube*, *center divider*) (Carrol & Conklin 2008). PMI is calculated as follows (Church & Hanks 1990):



Beginning on the right, the numerator corresponds to how often two words, such as *bad* and *apple*, are likely to co-occur. In other words, this is the probability of two words, (e.g., *bad* and *apple*) appearing together. The denominator factors out chance by multiplying the independent probabilities of each word appearing on its own. For example, for *bad apple*, it considers the likelihood of *bad* appearing on its own and multiplies this by the likelihood of *apple* appearing on its own. Thus, the denominator accounts for the possibility that two words might co-occur simply because they are both common, rather than because they have a meaningful association.

After calculating the probability ratio, the final step is to apply a logarithmic transformation to account for the exponential nature of probability ratios. Taking the logarithm compresses the range of values, preventing extreme differences from disproportionately influencing the measure while preserving meaningful distinctions between word associations. Additionally, the log transformation converts PMI into an additive, rather than multiplicative measure, which allows for direct comparison across word pairs. Finally, logarithmic transformation centers the PMI scale around zero, where a value of zero indicates that two words co-occur at a rate expected by chance, while positive and negative values reflect deviations

from this expectation, facilitating a more intuitive interpretation of co-occurrence strength.

2. Alternative measures of phrasal formulaicity

While it was deemed the most appropriate for this work, PMI is not the only measure of phrasal formulaicity. Other widely used measures include mutual information (MI), t-scores, and cloze probability, the latter of which is significantly more common in psycholinguistic work. Subsections 2.1-2.3 introduce these alternative measures, comparing them with PMI and further highlighting the motivation for the selection of this metric.

2.1. Mutual information

Mutual information (MI) is a closely related alternative method for quantifying co-occurrence and association strength. MI is defined as "the reduction in uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another." (Manning & Schuetze 2000:66). Thus, it measures the overall dependency between two sets of words. This is done by considering all combinations of the words' occurrences and non-occurrences.

The key difference between MI and PMI, which is derived from MI, can be seen in the equation below. Unlike PMI, which isolates the association strength of a specific phrase, MI considers not only how often two words co-occur but also the likelihood of each word occurring independently, as well as cases where neither word appears. For *bad apple,* this would include considerations of the probability of both words appearing together (p(*bad* + *apple*)), the probability of *bad* occurring without *apple* (p(*bad* + not *apple*)), the probability of *apple* occurring without *bad* (p(not *bad* +

apple)), and the probability of neither word appearing in the corpus p((not *bad*+not *apple*). This can be seen in the following equation:

$$I(X;\ Y) = \sum_{\{x,y\}} P(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

This means that MI incorporates instances where *bad* appears without *apple*, where *apple* appears without *bad*, and even where neither word appears at all. In contrast, PMI identifies the degree to which the observed co-occurrence of a phrase exceeds that which would be expected by chance, making it a more direct measure of lexical association strength. Essentially, PMI measures two specific points within a MI distribution while MI is the expected PMI over all possible outcomes.

While MI provides a more comprehensive measure of word dependency, it has largely been replaced by other metrics, particularly PMI, as MI is known to emphasize highly frequent words and very rare words (Kilgarriff & Kosem 2012). As a result, MI is less useful for identifying strong lexical associations in practice, as high MI scores may not necessarily indicate meaningful co-occurrence patterns. PMI improves upon the core concept of MI, refining it for linguistic applications as a more precise measure of lexical association strength.

2.2. T-scores

The second co-occurrence measure is the t-score. T-scores use individual word frequency to determine whether the co-occurrence is significantly different from what would be expected by chance. This is calculated as follows:

$$T - score = \frac{O_{xy} - E_{xy}}{\sqrt{O_{xy}}}$$

Here, O_{xy} is the observed frequency of the co-occurrence of words x and y. E_{xy} is the expected frequency of the co-occurrence of words x and y, assuming independence. This is calculated as follows, where $f(x)$ and $f(y)$ are the individual frequencies of the words x and y, and N is the total number of word pairs in the corpus.

$$E = \frac{f(x) * f(y)}{N}$$

Higher t-scores indicate a stronger association between words, suggesting they co-occur more often than would be expected by chance (are more formulaic). While t-scores may be useful in conjunction with a probability-based metric, such as PMI, there are a number of reasons why this measure was not selected.

T-scores incorporate elements of absolute frequency as part of their calculation. As a consequence, meaningful t-scores require high observed frequencies for accurate results. If a highly formulaic phrase has a low frequency in a given corpus, the t-score may underestimate the statistical or linguistic significance because reliability decreases with smaller sample sizes. Because PMI focuses on association strength rather than frequency, considering the likelihood of words appearing together relative to their independent probabilities, it is better suited to the detection of formulaic phrases with low frequencies.

A second problem with t-scores comes from an overemphasis of word pairs that include highly frequent words. Because t-scores prioritize co-occurrence frequency rather than association, pairs of words that frequently appear together due to high independent frequency or other influences, such as syntax, will receive higher scores. For example, the words *boy* and *that* have a high independent frequency.

They also commonly appear together due to the fact that *that* signals a clausal boundary, allowing one to elaborate on *boy*. While *boy that* is not a collocation, the t-score for this two-word pair would be relatively high. This is in spite of the fact that, because *that* can introduce a number of clauses (e.g., relative, compliment, extrapositional, comparative, etc.), it commonly appears with a number of other nouns in addition to *boy*. Thus, because t-scores are heavily influenced by absolute frequency, even modest deviations in co-occurrence between high-frequency words like *boy* and *that* can produce high t-scores. Conversely, the PMI value for *boy that* would be quite low. This is because PMI prioritizes association strength, factoring out the probability of highly frequent words appearing together by chance due to high individual frequency or due to predictable syntactic patterns.

Questions regarding the use of t-scores with idioms are not limited to its reliance on frequency. Because the metric makes use of normalization, it is less able to detect formulaic phrases or to model the degree of difference between relatively fixed formulaic phrases and word pairs that are not collocates. A final criticism of t-scores has been that "differences in average t-scores are more likely to be the result of how speakers combine words, and not of which words they retrieve" (Zimmerman et al. (2016:7). Thus, PMI is a more appropriate measure for psycholinguistic work.

2.3. Cloze probability

A popular alternative to objective measures is the subjective measure of cloze probability. Cloze is a measure of the relationship between a word and the context in which it appears and can be defined as the proportion of people who will choose a specific word to "close a gap" (fill in a blank in a completion task) in a given context (Taylor 1953, Kučera & Francis 1987, Coulson 2007, Block & Baldwin 2010). For example, (1) has high cloze probability because, for most native speakers, the word

mac immediately comes to mind. However, this same word has low cloze probability in (2).

(1) My favorite hamburger at McDonalds is a big ___.

(2) When we go out to eat, my favorite thing to order is a big ___.

Thus, cloze reflects the entropy of the final word in a given word string, making it an ideal metric for non-formulaic phrases as it provides a measure of the degree to which context impacts predictability calibrated to a specific community of practice. However, while often used in idiom work, it is not ideal. Due to their high degree of formulaicity, even in poorly supported context, the final word in an idiomatic phrase may be highly predictable because words within an idiomatic phrase provide unusually ***S***trong clues about upcoming words. Cloze is unable to account for two independent sources that shape word expectations and reduce surprisal. This problem is avoided by using an objective measure, such as PMI.

3. PMI calculation

The following section describes how the PMI values used in this work were obtained. This includes an overview of the script verification process, accomplished via output evaluation, as well as a discussion of considerations for future work.

3.1. PMI python script

To calculate PMI values, a 6-million-word corpus was created by scraping locations known for having written text reflective of informal speech, such as Yelp review posts, blogs, and movie or TV scripts. The text was cleaned during a preprocessing phase to normalize and format data. Examples of preprocessing steps included converting all characters to lowercase to ensure case insensitivity and performing tokenization using the Natural Language Toolkit (NLTK), which segmented the text into individual words while preserving meaningful linguistic units.

Additionally, predefined regular expressions and a punctuation filter were used to remove non-word characters and extraneous symbols from the tokenized text.

After preprocessing, the first step of the PMI script was to generate bigrams, or sequences of two consecutive words, using the ngrams function from NLTK. For example, a sentence such as *The bad apple fell* would produce the bigrams: *The bad*, *bad apple*, and *apple fell*. Each bigram was stored as a concatenated string to facilitate retrieval and processing. The resulting list of bigrams served as the primary input for PMI calculations.

Next, the individual probabilities of words and bigrams in the corpus were calculated by counting the number of occurrences of each word and dividing by the total number of words in the corpus:

$$P(w) = \frac{count(w)}{total\ number\ of\ words}$$

Similarly, bigram probabilities were determined by counting how often each bigram appeared and dividing by the total number of bigrams:

$$P(w_1, w_2) = \frac{count(w_1, w_2)}{total\ number\ of\ bigrams}$$

These calculations were pickled and retained for use in the next step, during which PMI values were computed using the formula for PMI stated in 3 above. Because the intended function of this program was not only to calculate PMI values for specific phrases but also to be able to return a list of phrases based on a specified

PMI value, this calculation was performed for all bigrams in the dataset. Thus, if a bigram appeared in the corpus, a PMI value was computed and stored[35].

Once PMI values were computed and stored for all bigrams in the corpus, these probabilities were used to compute the PMI values for specific word pairs and to identify other phrases with identical PMI values. During this final step, the list of 550 candidate phrases was entered and PMI values were printed. Additionally, for each of the 550 candidate phrases, a list of 50 phrases with a matching PMI value[36] were returned. This was accomplished using a lookup function that searched the previously computed and stored PMI values, retrieving those matching the PMI score for a specific phrase.

3.2. Evaluation and external comparison

To ensure that script-calculated PMI values aligned with linguistic expectations, externally calculated PMI values for the 550 candidate phrases were obtained from COCA and Sketch Engine[37] and compared against values returned by the PMI script. The goal of this comparison was to evaluate the relative ranking of phrases and to compare the "shape" of scores, not to determine whether values were the same between resources. Instead, the analysis focused on relative rankings of phrases across different corpora, ensuring that the distribution of high- and low-PMI phrases remained comparable.

[35] If a bigram occurred only once and its individual words were also infrequent, the returned value could be skewed, providing a less than reliable measure of association strength. For this reason, individual word frequency and phrasal frequency were taken into account when selecting phrases.
[36] Defined as PMI within .01 of a candidate phrase.
[37] Co-occurrence data for the most frequent word pairs was purchased from COCA. MI for missing phrases was calculated using the collocates search function on english-corpora.org. All Sketch Engine values were obtained by first performing a concordance search then using the advanced collocations options to return the three association strength scores. In both resources, the word window was set such that no intervening words were allowed.

Direct comparison between resources is not possible for two reasons. First, PMI values are inherently dependent on corpus composition. Available corpora of a sufficient size generally underrepresent linguistic patterns common to informal language, particularly informal spoken language. This fact partially motivated the creation of a new corpus for this project that prioritized the inclusion of informal items such as movie and TV scripts, blog posts, reviews, transcripts of dialogue such as interviews, etc.). Second, the exact equations used by the external resources differ slightly. While scores are generally accepted as comparable (Davies n.d), COCA uses a modified version of MI (1) that is more closely related to PMI than true MI as it is scaled according to both span and corpus size. This stands in opposition to traditional MI, which sums over all possible outcomes, such as all word pairs (Manning & Schultz 2000).

(1) COCA's MI equation (Davies n.d., https://www.english-corpora.org/mutualInformation.asp).

$$MI = log_2 \frac{f(x) * f(y) * N}{f(x) * f(y) * span}$$

The exact equation used by Sketch Engine is less clear. While the equations in (2)-(4) are provided in various resources, Kilgarriff et al. (2014) state that, as of 2006, MI has been calculated as a scaled version of logDice, a less common measure of association strength (Rychly 2008, Lexical Computing 2015, Kilgarriff et al. 2014, Kilgarriff 2020, Davies n.d., see also Evert 2008).

(2) Sketch Engine's MI equation (Rychly 2008).

$$MI = log_2 \frac{f(xy)N}{f(x)f(y)}$$

(3) Sketch Engine's MI3 equation (Rychly 2008).

$$MI^3 = log_2 \frac{f^3(xy)N}{f(x)f(y)}$$

(4) Sketch Engine's LogDice equation (Rychly 2008).

$$\frac{2 * f(xy)}{f(x) + f(y)}$$

Thus, fundamental differences between the PMI corpus and calculations used in this research, COCA, and Sketch engine, prevent direct comparison. However, the goal was not to assess the mathematical validity of PMI or MI but instead to ensure that the implementation of PMI in the script was correctly executed. To this end, returned PMI values were compared to the three types of available association scores: COCA's MI score, Sketch Engine's MI score, and Sketch Engine's MI3 score. Overall, the PMI script values were closely aligned with the external resources, particularly with COCA's MI score and Sketch Engine's logDice (see Table 14 for an example). To avoid potential inconsistencies, phrases that differed significantly between resources (e.g., high association values in one corpus but low association values in another) were excluded from further consideration as potential stimuli.

Resource	**Value**
PMI Script	5.393
COCA	4.87
Sketch Engine MI	8.03
Sketch Engine MI3	34.88
Sketch Engine logDice	5.49

Table 14. Comparison of word-association scores. This table provides the results for slim chance as an example of value similarities between resources.

3.3. Limitations and future considerations

The PMI script was developed specifically for this project. While it is a useful tool for calculating PMI reflective of informal language and quickly identifying two-word phrases with matching or specified values, there are some inherent limitations that should be considered. These include the treatment of missing and negative PMI values (3.4.1) and the operationalization of "co-occurrence" for bigrams (3.4.2), and word-length considerations (3.4.3).

3.3.1. Missing and infrequent bigrams

While a discussion of stimuli creation via PMI matching can be found in Appendix 1, two points must be mentioned here. First, if a word pair never co-occurred in the corpus, no bigram was formed. The PMI value was set to infinity to reflect the absence of co-occurrence. As a result, phrases on the candidate list not found in the corpus returned a PMI value of "undefined". Such phrases were removed and not considered further. Next, some observed bigrams received negative PMI values, indicating weaker-than-e*x*pected associations. Like phrases with an undefined value, phrases with a negative PMI value were also removed from the list of candidate test phrases as they did not reach the minimum PMI value required to be considered for inclusion.

Should this script be used for more general PMI calculation purposes, alternative treatment of non-existent and negative values should be considered. One option for handling negative values is to use positive pointwise mutual information (PPMI). A commonly used PMI alternative, PPMI neutralizes negative PMI values by converting them to 0. This is done with a selection equation specifying that PMI

values above 0 should be retained while those with a negative value should be replaced by 0. This is expressed as follows:

$$PPMI(x,y) = max(log_2 \frac{P(x,y)}{P(x)P(y)}, 0)$$

While negative PMI calculations might provide insights into weak or absent associations, it is a computationally expensive process that was not relevant to the goals of this study, which analyzed observed phrases to understand their formulaicity and association strength. Despite optimization of script efficiency, given the size of the corpus, calculating PMI for all possible combinations would put an immense strain on available RAM and may not have been possible on a personal laptop. More importantly, an extremely large sample is needed for negative values to be reliable. Thus, should negative values have been calculated, they would have been unreliable. Interestingly, this is mirrored in human judgements – while native speakers are able to evaluate relatedness, Jurafsky & Martin (2024) note that it is unclear whether it is possible for native speakers to reliably evaluate unrelatedness.

There are some implications from this decision. By excluding unobserved pairs, we are unable to gain insight into which words could be statistically "avoiding" each other, or which word pairs may be candidates for serving as examples of phrases that co-occur less often than would be expected. Additionally, because this approach inherently focuses on positive associations, any descriptive statistics regarding the overall nature of co-occurrence within the corpus would exhibit a positive skew. Thus, because PMI was calculated only for observed pairs, broader conclusions

about the overall structure or statistical properties of co-occurrence within the corpus could not be reliably drawn.

3.3.2. Bigram co-occurrence probability

There are two common methods for calculating co-occurrence probability: the sliding window approach and the direct adjacency approach. In this work, PMI was calculated using direct adjacency rather than a sliding window. While this methodology is valid for two-word, relatively fixed idioms, it has implications for phrases with some degree of flexibility, particularly those longer than two words.

In a sliding window approach, a window size within which two words are considered to be co-occurring is defined. For example, a sliding window of four words considers two words that appear within four words of each other as co-occurring, regardless of the intervening items. Thus, in a sentence such as *The bad apple fell,* not only would *bad apple* be considered a valid co-occurrence, but also pairs including *the apple, the fell,* and *bad fell.* This method is useful for capturing more flexible associations, particularly in languages with free word order or in cases where intervening words frequently appear. For example, in *the apple that was bad*, *bad apple* is counted as a co-occurring pair because the words appear within the four-word window.

By contrast, a direct adjacency approach requires that two words occur sequentially, with no intervening items. For instance, in the sentence *The bad apple fell,* only *bad apple* would be considered a valid co-occurrence under direct adjacency, whereas the separated words *bad* and *fell* would not. This approach is stricter than a sliding window and is particularly suited to fixed phrases, where word order and immediate proximity are critical. Additionally, idiomatic phrases often rely on the strict juxtaposition of specific words to retain idiomatic meaning (e.g., *spill beans*

does not have the same meaning as *spill the beans*). However, not all two-word phrases are fully fixed. This can be problematic as, under a direct adjacency approach, *bad apple* in the passive phrase *the apple that was bad* would not be counted as a co-occurring pair because the words are not adjacent.

In this study, the direct adjacency approach was used to calculate PMI. This choice was motivated by the focus on two-word phrases, which inherently limit flexibility. While the role of flexibility should be considered in future work, it was outside the scope of this dissertation. Further motivation came from the need to match idiomatic phrases with non-idiomatic collocations based on PMI values in a manner able to address claims of holistic representation. Because phrases vary in their degree of allowable flexibility, using a sliding window could result in matching a highly fixed phrase with a more flexible one. While both might have similar PMI values, the flexible phrase may not exhibit the same conventionalized adjacency. According to a holistic approach, this may alter how the phrase is mentally represented and processed – a fixed idiomatic phrase may be holistically represented while a more flexible phrase may not be.

3.3.3. Multiword mutual information

While less common in computational and psycholinguistics, multivariate pointwise mutual information (MPMI) must be mentioned for the purposes of methodology and generalizability. PMI is intended for use with two-word phrases, in large part motivating the use of two-word idioms in this work. Two-word idioms favor the form of modifier + noun, and they are less flexible overall than longer idioms. Generally speaking, within psycholinguistics larger effects have been observed for longer idioms, such that there may be an association between length and the expected amplitude of experimental impact, particularly for studies investigating prediction (Ellis et al. 2009, Carrol 2015, Zheng et al. 2022). From this perspective, the

findings observed in this work show particular promise for extension to phrases of different lengths and forms. Prior to such an extension, the PMI equation would need to be modified to account for additional words.

Accounting for additional words is not as simple as adding an extra term (e.g., w_z) to both the numerator and denominator of the equation. This is because extending PMI to three or more words introduces a non-linear effect that distorts the results. The joint probability of w_1, w_2, …, w_n, which would be represented by the numerator rapidly decreases as N increases, since the probability of all words appearing together becomes increasingly rare. At the same time, the denominator increases rapidly. This is because the product, $p(w_1)$, $p(w_2)$, …, $p(w_n)$, reflects the probability of each word occurring independently. As more words are included, the resulting value grows disproportionately large relative to their actual joint probability. As a result, PMI values become more extreme. Further, joint probability decreases much faster than independent probability increases, further contributing to value distortions leading to artificially large positive or negative values. Instead, one must account for how each item relates to all the others, both individuall$y$ and in combination. MPMI offers such a modification, which can be implemented in a variety of ways. Two implementation options as well as a discussion of considerations for selecting the most appropriate modification for a specific use can be found in Van de Cruys (2011).

APPENDIX II

STIMULI CREATION: PMI-MATCHED PHRASE SELECTION

This appendix walks through the PMI pairing process for idioms and non-idioms using the phrase *pink slip* as an example.

1. Identifying matches from calculated PMI

The PMI value of each phrase was calculated based on co-occurrence frequency in our corpus (see Appendix I for information regarding PMI calculation). The PMI value for pink slip was 10.035. Using a python script that calculated the PMI values of all bigrams in the corpus, a reverse search for phrases other than pink slip with a PMI value of 10.035 was conducted and the first 50 PMI-matched phrases were returned. Due to the nature of such a corpus, returned phrases could be literal or figurative and were syntactically unrestricted. Thus, an over abundance of PMI-matched phrases were returned to ensure the identification of an ideal non-idiomatic collocation as only non-idiomatic, nominally headed phrases could serve as non-idiomatic collocation stimuli.

Phrase	**PMI value**	**Matches**
pink slip	10.035	water heater
		fossil fuel
		mr. smith

		racial biases
		operation fallout
		openly hostile
		jobs repairing
		distributed throughout
		turned inward
		mainstream democrat
		screen adaptation
		stylized games
		st. peter
		highly commendable
		web developers
		investment guru
		arguments ibm
		born hungry
		twelve paces
		meatball surgeon
		short closings
		help joey
		tiny clusters
		buying yachts
		complete exclusivity
		hamas flip

		slutty blonde
		cabotage travel
		wax mustache
		overly ambitious
		relentless march
		ideological breakdown
		tommy pinto
		cardiothoracic spot
		wooden owl
		aging outright
		longitude fifty
		shouted lyrics
		connors intimate
		sleep nice-nice
		fouls citizens
		integer thats
		precariously short
		your soul
		russell hitchcock
		screen magnify
		encore presentation
		falsely modest
		overweight ranging

		soul unburden
		contested unlike

2. Removal of returned phrases that were not nominally headed

The majority of returned candidate phrases were nominally headed. However, because returned phrases were unrestricted, returned items could include prepositions, adverbs, verbs, determiners, etc. In the case of pink slip, the following phrases were removed.

Removed due to part of speech
openly hostile
jobs repairing
distributed throughout
turned inward
highly commendable
born hungry
help joey
buying yachts
complete exclusivity
overly ambitious
aging outright
connors intimate
sleep nice-nice

fouls citizens
integer thats
precariously short
screen magnify
falsely modest
overweight ranging
soul unburden
contested unlike

3. Phrase-specific considerations

Candidates were further curated by removing phrases containing proper nouns or pronouns, phrases containing words not widely known, and incomplete phrases.

Removed	Reason
mr. smith	Proper noun
stylized games	Not well known/known to a specific community of practice
st. peter	Proper noun
arguments ibm	Proper noun, incomplete phrase
meatball surgeon	Not well known/known to a specific community of practice
hamas flip	Proper noun, incomplete
short closings	Not well known/known to a specific community of practice
cabotage travel	Not well known/known to a specific community of practice
tommy pinto	Proper noun
cardiothoracic spot	Not well known/known to a specific community of practice
longitude fifty	Incomplete, not well known/known to a specific community of practice
your soul	Pronoun
russell hitchcock	Proper noun

4. Optimal candidates

Remaining candidates were further narrowed by considering word and phrase length. The ideal match should be as orthographically similar to its idiomatic counterpart as possible. In this case, the ideal match would have a 4-character first word and a 4-character second word. In many cases, an exact match did not exist. This was addressed by balancing length across all stimuli.

Remaining options	Length
water heater	5 + 6
fossil fuel	6 + 4
racial biases	6 + 6
operation fallout	9 + 7
mainstream democrat	10 + 8
screen adaptation	6 + 10
web developers	3 + 10
investment guru	10 + 4
twelve paces	6 + 5
tiny clusters	4 + 8
slutty blonde	6 + 6
wax mustache	3 + 8
relentless march	10 + 5
ideological breakdown	11 + 9
wooden owl	6 + 3
shouted lyrics	7 + 6

encore presentation	6 + 12

5. Final candidate selection

In cases where more than one candidate phrase remained, a winner was chosen based on a number of practical considerations. These included whether there was any chance that a candidate phrase might have a negative or socially charged connotation and whether it was common for the phrase to appear in the sentence-final position of a given sentence.

APPENDIX III

FACTORIAL DESIGN

The following chart provides a visual description of the factorial stimuli design.

Idiom				Non-Idiomatic Collocations				Filler
DIS	**PL**	**DM**	**PMI**	**DIS**	**PM**	**DM**	**PMI**	**Type**
Y	Y	Y	PMI-H	N	N	Y	PMI-H	IP
Y	Y	Y	PMI-H	N	N	Y	PMI-H	NIP
Y	Y	Y	PMI-M	N	N	Y	PMI-M	IP
Y	Y	Y	PMI-M	N	N	Y	PMI-M	NIP
Y	Y	Y	PMI-L	N	N	Y	PMI-L	IP
Y	Y	Y	PMI-L	N	N	Y	PMI-L	NIP
Y	Y	N	PMI-H	N	N	N	PMI-H	IP
Y	Y	N	PMI-H	N	N	N	PMI-H	NIP
Y	Y	N	PMI-M	N	N	N	PMI-M	IP
Y	Y	N	PMI-M	N	N	N	PMI-M	NIP
Y	Y	N	PMI-L	N	N	N	PMI-L	IP
Y	Y	N	PMI-L	N	N	N	PMI-L	NIP
Y	N	Y	PMI-H	N	N	Y	PMI-H	IP
Y	N	Y	PMI-H	N	N	Y	PMI-H	NIP
Y	N	Y	PMI-M	N	N	Y	PMI-M	IP

Y	N	Y	PMI-M	N	N	Y	PMI-M	NIP
Y	N	Y	PMI-L	N	N	Y	PMI-L	IP
Y	N	Y	PMI-L	N	N	Y	PMI-L	NIP
Y	N	N	PMI-H	N	N	N	PMI-H	IP
Y	N	N	PMI-H	N	N	N	PMI-H	NIP
Y	N	N	PMI-M	N	N	N	PMI-M	IP
Y	N	N	PMI-M	N	N	N	PMI-M	NIP
Y	N	N	PMI-L	N	N	N	PMI-L	IP
Y	N	N	PMI-L	N	N	N	PMI-L	NIP
N	Y	Y	PMI-H	N	N	Y	PMI-H	IP
N	Y	Y	PMI-H	N	N	Y	PMI-H	NIP
N	Y	Y	PMI-M	N	N	Y	PMI-M	IP
N	Y	Y	PMI-M	N	N	Y	PMI-M	NIP
N	Y	Y	PMI-L	N	N	Y	PMI-L	IP
N	Y	Y	PMI-L	N	N	Y	PMI-L	NIP
N	Y	N	PMI-H	N	N	N	PMI-H	IP
N	Y	N	PMI-H	N	N	N	PMI-H	NIP
N	Y	N	PMI-M	N	N	N	PMI-M	IP
N	Y	N	PMI-M	N	N	N	PMI-M	NIP
N	Y	N	PMI-L	N	N	N	PMI-L	IP
N	Y	N	PMI-L	N	N	N	PMI-L	NIP
N	N	Y	PMI-H	N	N	Y	PMI-H	IP
N	N	Y	PMI-H	N	N	Y	PMI-H	NIP

N	N	Y	PMI-M	N	N	Y	PMI-M	IP
N	N	Y	PMI-M	N	N	Y	PMI-M	NIP
N	N	Y	PMI-L	N	N	Y	PMI-L	IP
N	N	Y	PMI-L	N	N	Y	PMI-L	NIP
N	N	N	PMI-H	N	N	N	PMI-H	IP
N	N	N	PMI-H	N	N	N	PMI-H	NIP
N	N	N	PMI-M	N	N	N	PMI-M	IP
N	N	N	PMI-M	N	N	N	PMI-M	NIP
N	N	N	PMI-L	N	N	N	PMI-L	IP
N	N	N	PMI-L	N	N	N	PMI-L	NIP

Key

Y: Yes, the phrase has this property

N: No, the phrase does not have this property

PMI-H: PMI value greater than 6

PMI-M: PMI value of 3-5.9

PMI-L: PMI value less than 2.9

IP: filler was created based on an idiomatic phrase

NIP: filler was created based on a non-idiomatic phrase

APPENDIX IV

RATINGS TASK SAMPLE

Appendix IV contains an example from the practice section of each survey. This page was presented after the instructions and provided a brief reminder of the response options. Full instructions were also provided at the bottom of the screen. After completing a practice block, participants advanced to a screen providing an explanation of expected responses.

1. Phrase type recognition task practice block

Response options

- Not idiomatic = the meaning of each word is added together to create the meaning of the phrase
- Unsure = cannot determine phrase type. Please use this only when absolutely necessary.
- Idiomatic = the meaning of the phrase is different than the meaning of each word added together

Practice

Please read the sentences below and evaluate the meaning that each phrase in bold has in the provided context. Select the response option that best fits your interpretation of each phrase. To see examples of phrases that are "not idiomatic", "unsure", or "idiomatic", hover over these word at the top of the response matrix. The full instructions have also been provided at the bottom of this screen as a reminder.

	Not Idiomatic	Unsure	Idiomatic
Often talking his friends into breaking rules, he was known to be a **bad apple**.	○	○	○
After biting a neighbor, I took him to a **dog trainer.**	○	○	○
Every Friday, the band held an informal **jam session**.	○	○	○
At Elias National Park, you can walk right up to the glaciers to experience them **first hand**.	○	○	○

2. Distributed idiomatic meaning task practice block

Practice 1

Consider the meaning that each phrase in bold has in the provided context. Determine whether or not every word can be assigned a portion of the phrase's meaning. Remember that a phrase is additive only if you can distribute the entire meaning of the phrase across the parts. If one word seems associated with a part of the phrase's meaning but the other word isn't associated with the remainder of the phrase's meaning or, if neither word is associated with the phrase's meaning, it is non-additive.

Response options

- Additive meaning = each word in the phrase is individually associated with a part of the meaning of the whole phrase
- Non-additive meaning = each word in the phrase is not individually associated with a part of the meaning of the whole phrase

	Additive meaning	Non-additive meaning
I attended my high school reunion in hopes of running into my **old flame**.	○	○
To get the oven clean, it's going to take some serious **elbow grease**.	○	○
Even more than the NFL, I love to watch **college football**.	○	○
For many young couples, the marriage conversation is a **hot potato**.	○	○

3. Partial literality task practice block

Practice

Please read the sentences below and evaluate the meaning that each phrase in bold has in the provided context. Determine whether the meaning that each word has outside of the phrase matches its meaning in the phrase. Select the response option that best fits your interpretation of each phrase.

Response options

- Both typical = both words in a phrase contribute their general meaning (that is the meaning that each word usually has outside of the phrase)
- One typical = one word contributes its general meaning and one word does not
- No typical = no words in the phrase contribute their general meaning

	Both typical	One typical	No typical
Even more than the NFL, I love to watch **college football**.	○	○	○
I only saw her for a **split second**.	○	○	○
When something bad happens, they always try to find the **silver lining**.	○	○	○
After he bit a neighbor, I took him to a **dog trainer**.	○	○	○

4. Dual meaning task practice block

Practice

Read each sentence and decide whether the phrase in bold is used literally or non-literally then determine whether that phrase could plausibly be used the other way.

Response options

- Literally only = The phrase has literal meaning and can be used literally only.
- Non-literally only = The phrase has nonliteral meaning and can be used non-literally only.
- Both literally and non-literally = The phrase has literal or nonliteral meaning and can be used literally and non-literally

	Literally only	Non-literally only	Both literally and non-literally
When the oven timer sounds, use a fork to check the **hot potato**.	○	○	○
If I mess up this speech, I'll be a **laughing stock!**	○	○	○
Losing the election left us with the bitter taste of **sour grapes**.	○	○	○
Even more than the NFL, I love to watch **college football**.	○	○	○