Wierzba, Marta & Brown, J. M. M. & Fanselow, Gisbert. 2023. Sources of variability in the syntactic flexibility of idioms. *Glossa: a journal of general linguistics* 8(1). pp. 1–41. DOI: https://doi.org/10.16995/glossa.8502

Sources of variability in the syntactic flexibility of idioms

Marta Wierzba, University of Potsdam, Germany, marta.wierzba@mailbox.org

J. M. M. Brown, University of Potsdam, Germany; Université de Lausanne, Switzerland, jessica.brown@unil.ch

Gisbert Fanselow†, University of Potsdam, Germany

Idiomatic verb phrases (e.g., *kick the bucket*, fig. 'to die') vary in their syntactic flexibility: they can undergo operations like, e.g., passivization ("*The bucket was kicked*") to varying degrees. We (re-)consider potential sources of this variability. It has been proposed that *compositionality* influences syntactic flexibility of idioms. In the first part of the paper, we reassess this finding from a methodological perspective by replicating earlier experiments on German and English, in which we change the previously used – and potentially biased – methods of measuring compositionality. Our results for German are compatible with the view that higher compositionality makes some of the tested structures more acceptable (most consistently: scrambling, prefield fronting, and *which*-questions), while we do not find a connection between compositionality and flexibility for English. In the second part of the paper, we present an additional experiment following up on the German findings. We extend the empirical domain and explore factors which – in contrast to compositionality – have the potential of explaining the syntactic flexibility of both idioms and non-idioms. We find that *definiteness* influences the flexibility of idioms and non-idioms in similar ways, supporting the view that both types of expressions are subject to the same grammatical rules. We discuss *referentiality* as a potential underlying semantic source for the behavior of both idioms and non-idioms.

1. Introduction

1.1 Syntactic flexibility of idioms and non-idioms

It is a consistent finding that on average, idioms have less syntactic flexibility than non-idioms. For example, while a non-idiomatic verb phrase like *eat the apple* can undergo passivization or nominalization without losing its meaning, this is less acceptable for an idiom like *kick the bucket* (fig. 'to die').

(1) a. Mary ate the apple.
b. The apple was eaten by Mary.
c. I want to talk about Mary's eating the apple.

(2) a. Mary kicked the bucket.
b. *The bucket was kicked by Mary.
c. *I want to talk about Mary's kicking the bucket.

That idioms are syntactically less flexible than non-idioms (i.e., less acceptable in non-canonical structures) on average has been observed based on intuitive judgments in the theoretical literature (e.g., Fraser 1970; Nunberg et al. 1994) and confirmed by experiments (see Section 2.2). The finding is illustrated schematically in **Figure 1** (left plot). There is also variability within idioms: some idioms are as syntactically flexible as non-idioms, while others are inflexible (**Figure 1**, right plot).

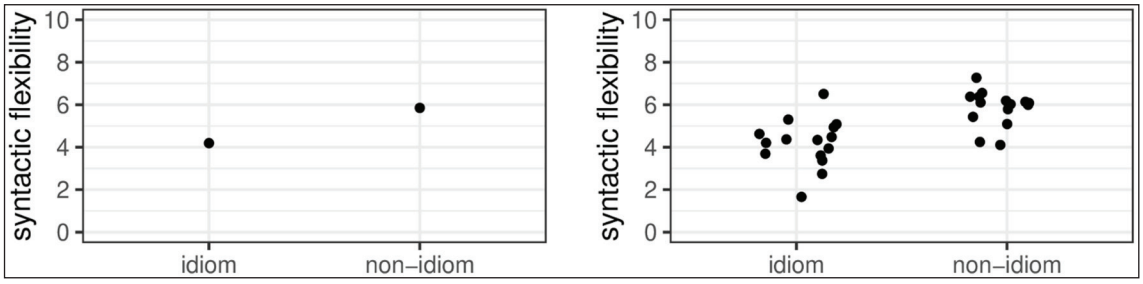

Figure 1: Illustration of means and distribution of idioms and non-idioms (hypothetical data).

In this paper, we address the following empirical questions concerning these observations: Qi – Why are some idioms more flexible than other idioms? Qii – Why are some non-idioms more flexible than other non-idioms? Qiii – Is there indeed a gap in syntactic flexibility between idioms and non-idioms, and if so, why?

By addressing Qi–iii, we aim to evaluate to what extent we can replace descriptively adequate, but stipulative rules as in (3i–iii) by more principled explanations as in (4i–iii) – ideally based on a linguistic property P as a common cause – which would be a step towards a grammar without

separate sets of grammatical rules for idioms and non-idioms and a deeper understanding of syntactic flexibility.

(3) Descriptively adequate answers to Qi–iii:
  (i) Idioms are ordered in their syntactic flexibility in a specific way:
  Idiom set A is more flexible than idiom set B, which is more flexible than…
  (ii) Non-idioms are ordered in their syntactic flexibility in a specific way:
  Non-idiom set A is more flexible than non-idiom set B, which is more flexible than…
  (iii) Idioms are syntactically less flexible than non-idioms.

(4) Goal: More principled/explanatory answers to Qi–iii:
  (i) Idioms that are highly *[P]* are syntactically more flexible (because…).
  (ii) Non-idioms that are highly *[P]* are syntactically more flexible (because…).
  (iii) Idioms are syntactically less flexible than non-idioms because they are typically less *[P]*.

1.2 Potential sources of variability

As a first step of describing the variability in syntactic flexibility among idioms – addressing question Qi – Fraser (1970) proposed a hierarchy ordering categories of idioms with respect to their syntactic flexibility, along the lines of (3i). According to this hierarchy, there is a set of idioms A containing, e.g., {bring down the house, put on a good face, …}, which are generally more syntactically flexible – compatible with a wider range of syntactic structures – than the idioms in a set B containing, e.g., {kick the bucket, shoot the bull, …}, etc.

In later work, researchers took steps towards a more principled answer to question Qi. A prominently discussed potential systematic source of syntactic flexibility in idioms is *compositionality*. Nunberg et al. (1994) suggested that idioms might be flexible if each of their parts makes an individual contribution to the figurative meaning (e.g., for *spill the beans,* fig. 'to reveal a secret': *spill* = 'reveal', *beans* = 'secret') and inflexible if they have a holistic figurative meaning that cannot be divided up in this way (e.g., *kick the bucket*). This provides a more principled answer to Qi, along the lines of (4i): the individual syntactic flexibility of an idiom would be systematically linked to a property P, and that property would be compositionality.

The compositionality hypothesis has been tested using corpus-based and psycholinguistic measures of compositionality, with varying results. The first goal of the novel experiments presented here is to revisit the compositionality hypothesis methodologically: we employ a different psycholinguistic measure of compositionality, which, we believe, avoids a potential confound present in previous studies. Our results confirm that compositionality is a relevant factor for syntactic flexibility at least in German, and does not affect all syntactic structures alike.

After reinforcing the robustness of this finding, a further goal of our paper is to revisit the compositionality hypothesis from a conceptual perspective. We argue that compositionality is

not sufficient as an explanation for the variability of idioms and non-idioms. Explaining the varying degrees of syntactic flexibility in idioms in terms of compositionality has the advantage of potentially providing answers to two of our questions, namely Qi and Qiii. First, a compositionality approach suggests that the reason why some idioms are more flexible than others (Qi) is because some idioms are more compositional than others. Second, a compositionality approach suggests that the reason why there is a gap in syntactic flexibility between idioms and non-idioms (Qiii) is because non-idioms are compositional, while not all idioms are. On the other hand, the question of why some non-idioms are more flexible than other non-idioms (Qii), which has received less attention so far in previous idiom studies, remains unanswered. Non-idiomatic expressions cannot vary in compositionality. Thus, any systematic variation in syntactic flexibility within non-idioms cannot be captured by compositionality.

This motivates investigating factors that could potentially influence the syntactic flexibility of both idioms and non-idioms and thus provide a unified account. This reasoning is in line with recent proposals that have stressed the importance of similarities in the behavior of idioms and non-idioms. Our experiments identify *definiteness* as a factor that has a similar influence on idioms and non-idioms; in contrast to compositionality, it can thus account for a part of the variability in syntactic flexibility in both types of expressions. However, definiteness cannot account for the general acceptability difference between idioms and non-idioms.

In our view, *referentiality* has the potential of providing a unified explanation for all observations mentioned above (idiom/non-idiom gap, variability in idioms, variability in non-idioms); in other words, it could be the missing property P in (4). Some aspects of our data are compatible with the idea of referentiality as a common source of syntactic flexibility of both idioms and non-idioms: we identify a class of non-idioms that are relatively inflexible, which might plausibly stem from lower referentiality; and variability in referentiality also provides a plausible explanation for the observed behavior of both idioms and non-idioms with respect to German scrambling. On the other hand, our data on pronominalization do not fully align with an explanation in terms of referentiality. We will discuss ways in which this question could be pursued further.

1.3 Structure of the paper

In Section 2, we provide background on the notions of idiomaticity and compositionality, and we discuss previous research. In Section 3, we present Experiments 1–2. They are replications of previous experiments of ours on German and English, which we reassess with an adjusted compositionality measure. Our results for German confirm the robustness of the earlier finding that compositionality influences syntactic flexibility of idioms. This provides the basis for a follow-up experiment on German presented in Section 4. We extend the data set and include further factors beyond compositionality. We discuss the influence of definiteness and referentiality on idioms and non-idioms, and how these factors interact with compositionality.

2. Background

2.1 Theoretical notions

In 2.2.1–2.1.2, we discuss the core notions that are relevant for Experiments 1–2 (idiomaticity, compositionality). Additional concepts relevant for Experiment 3 (definiteness, referentiality) will be discussed in Section 4.

2.1.1 Idiomaticity

Idioms are often seen as a "fuzzy category" (Nunberg et al. 1994:492) that is not easy to define precisely. Prototypical examples of idioms like *kick the bucket* or *spill the beans* share the characteristic that the meaning and/or function of the whole expression is conventionalized, i.e., it is not fully derivable from knowledge about its parts (Nunberg et al. 1994, ibid.). For example, a speaker who knows the meaning and use of *kick* and *bucket* in isolation but is not familiar with the idiom *kick the bucket* could not predict that the expression is conventionally associated with the meaning *to die*. A similar concept is that of "constructions" as expressions whose "form or function is not strictly predictable from its component parts" (Goldberg 2003:219). However, not all expressions that are "conventionalized expressions" or "constructions" in this sense would standardly be referred to as idioms. Idioms also prototypically involve some kind of figurative/metaphorical speech: e.g., collocations with a specific, conventionalized, but non-figurative meaning like the fixed expression *tax and spend* (referring to a certain government policy; Nunberg et al. 1994:494) are usually not categorized as idioms; but the distinction is often not clear-cut.

The fuzzy and multi-dimensional nature of idioms makes it difficult to conceptualize idiomaticity as a formal property in the grammar: it is not straightforward what formal properties a [+/− idiomatic] feature would correspond to. Thus, if there is indeed an observable difference in syntactic flexibility between prototypical idioms like *kick the bucket* and non-idioms like *eat the apple*, an explanation in terms of more precisely definable linguistic properties is called for.

2.1.2 Compositionality

We take a complex expression to be (semantically) compositional if its meaning can be derived from the meaning of the individual parts and general semantic mechanisms, like combining functions and arguments (going back to ideas by Frege 1891; formalized, a.o., by Montague 1974). We follow Nunberg et al.'s (1994) view that all idioms are highly conventionalized, but some of them can nevertheless be compositional with respect to their figurative meaning.[1] To illustrate this idea, let us elaborate more on the examples *spill the beans* vs. *kick the bucket*. Both are

[1] Nunberg et al. (1994) use the term "decomposable" instead of "compositional"; we use latter term here, but mean the same notion.

conventionalized: a speaker who knows the meaning of *spill* and *beans* does not necessarily know the idiomatic meaning of *spill the beans*; the same holds for *kick the bucket*. However, a speaker who *is* familiar with the idiom *spill the beans* might be able to establish a relation between the literal and the figurative meaning of the idiom in a way not available for *kick the bucket*. In *spill the beans*, *spill* can be understood as figuratively corresponding to the concept of *revealing*, and *the beans* to the *secret*; and these individual figurative meanings then combine compositionally to the complex meaning *reveal the secret*. Thus, in a sense, both the verb and the object have an individual (figurative) meaning within the idiom here. For k*ick the bucket*, finding such individual figurative meanings for *kick* and *the bucket* that would then combine compositionally is less straightforward.

However, if the compositionality of an idiom is conceived of as a native speaker's intuitive ability to assign individual figurative meanings, compositionality is subjective: possibly there are speakers who do not perceive *spill the beans* as compositional, and, conversely, also speakers who see *kick the bucket* as compositional (e.g., by having *kick* correspond to *end*, and *the bucket* to *life*). We will treat compositionality as a subjective and gradient property here: in Experiments 1 and 2, we collect individual speakers' judgments and derive a gradient compositionality measure for idioms.

2.2 Previous research

In Section 2.2.1, we summarize previous experimental findings concerning the effect of compositionality on syntactic flexibility. Section 2.2.2 discusses previously used methods of measuring compositionality. In Section 2.2.3, we discuss how the present paper builds on and differs from previous research. Section 2.2.4 provides background on research on parallels between idioms and non-idioms, which we aim to contribute to.[2]

2.2.1 Previous findings on idiom compositionality and syntactic flexibility

Gibbs & Nayak (1989) collected compositionality judgments and tested syntactic flexibility (as well as other idiom properties, a.o., transparency); see 2.2.2. for their method. They report an effect of compositionality on syntactic flexibility and conclude that compositional idioms are "more syntactically flexible" (Gibbs & Nayak 1989:100) than non-compositional ones with respect to some of the tested structures, in particular adjective insertion and passive. They report that compositionality also significantly affected the participants' reactions to pronominalization of a part of the idiom.

Tabossi et al. (2008) replicated the study for Italian. They also report an effect of compositionality on syntactic flexibility, but only for adverb insertion, while no contrast was found for any of the other tested structures.

[2] Our focus is on psycho-linguistic experiments and corpus examples here; see, a.o., Lebani et al. (2015) and Wulff (2009) for a review of methods from the domain of computational linguistics and how they relate to human compositionality judgments.

In three previous experiments, we investigated the syntactic flexibility of German and English idiomatic verb phrases (Wierzba et al. 2023). The following German structures were tested in these experiments: object movement to the prefield (the pre-verbal position in V2 clauses), object left dislocation, object scrambling, object pronominalization, passive, nominalization, object *which*-question. The following English structures were tested: passive, pronominalization, a cleft-like construction ("Kicking the bucket is something that…"), and two types of nominalization.

We found a robust effect of compositionality on syntactic flexibility in German: non-compositional idioms consistently showed lower acceptability in non-canonical syntactic constructions than compositional ones, while they were similarly acceptable in sentences with canonical word order.

For English, the picture was different: a gap between non-compositional and compositional idioms was observed as well, however, it was not significantly larger in the non-canonical structures than in the canonical baseline. Thus, in contrast to German, the gap could not be attributed to limited syntactic flexibility of non-compositional idioms in English.

In both languages, compositional idioms were as acceptable as non-idioms in most of the non-canonical syntactic structures, but some deviated from that pattern: for *which*-questions in German and the cleft-like construction in English, a larger part of our idiom set (including some of those that had been categorized as compositional) was judged as degraded than in the other constructions. This result was in line with the theoretical assumption that these two syntactic constructions impose a semantic requirement on the object and are thus less compatible with idiomatic VPs (where the object does not necessarily have an individual meaning, in particular when it is perceived as non-compositional, see Section 2.1.2) than constructions without such a requirement.

2.2.2 Previously used methods

To estimate the syntactic flexibility of idioms, Gibbs & Nayak (1989) asked participants to rate how similar the meaning of sentence pairs like (5) was. One sentence contained an idiom (*lay down the law*) and the other a non-idiomatic paraphrase (*give strict orders*). The sentence pairs were presented in various syntactic variations, a.o., structures containing passivization as in (5), or nominalization. High similarity ratings across all tested structures were interpreted as an indicator of 'syntactically flexible' idioms, whose idiomatic reading remains intact even with modified syntax.

(5) a. The law will be laid down when Jane's boyfriend finds out where she's been.
b. Strict orders will be given when Jane's boyfriend finds out where she's been.

The use of idiom-paraphrase pairs to assess syntactic flexibility has been adopted by later studies (e.g., Abel 2003, Tabossi et al. 2008), but criticized by Maher (2013) and Wierzba et al. (2023) – one problem is that similarity judgments of pairs like (5) can depend not only on the

passivizability of the idiom, but also of the paraphrase. We follow the latter approaches in using acceptability ratings instead of similarity ratings to estimate the syntactic flexibility of idioms.

Gibbs & Nayak (1989) also used idiom-paraphrase pairs to estimate idiom compositionality. They presented 40 idiom-paraphrase pairs similar to (5) and asked the participants "to decide whether the individual words in each expression made some unique contribution to the phrase's nonliteral interpretations" (Gibbs & Nayak 1989:108). For example, in (5), participants judged whether the verb *lay down* and the object *the law* each make an individual contribution to the idiom's figurative meaning of *giving strict orders*. If a participant said no, the idiom was categorized as "non-decomposable"; if they said yes, the idiom was categorized as "decomposable", and the participants were asked to further judge whether the relation between the literal and figurative meaning was direct ("normally decomposable") or indirect ("abnormally decomposable"); the latter more fine-grained subcategorization will not be relevant here.

Methodological criticism was raised by Maher (2013): all idioms tested by Gibbs & Nayak (1989) consisted of a transitive verb and an object, but this was not the case for the paraphrases. While some of them also involved a transitive verb and an object in parallel to the idiom (*lay down the law/give strict orders)*, others only consisted of an intransitive verb (*kick the bucket/die*), an intransitive verb and an adjunct (*chew the fat/talk aimlessly),* or a predicative construction *(pack a punch/to be powerful)*. All idioms whose paraphrase deviated from the idiom's transitive *verb + object* structure were categorized as "non-decomposable" by the majority of the participants. Maher (2013) hypothesized that the form of the paraphrase might have influenced participants' reactions in the categorization task: it could be easier to draw a connection between an idiom's individual literal parts and its figurative meaning if the idiom and the paraphrase have parallel structures.

In our previous experiment on English, which examined whether compositionality affects syntactic flexibility, the categorization was adopted from Gibbs & Nayak's (1989) empirically based lists of "decomposable" and "non-decomposable" idioms. In our previous experiment on German, we intuitively categorized the idioms as "compositional" or "non-compositional". In this paper, we will replicate these studies, but will use a method to estimate the compositionality of an idiom neither requiring introspective categorization by the authors nor choosing literal paraphrases. Thus, we are going to reassess the previous findings with a new empirically grounded measure.

2.2.3 Demarcating the scope of this paper

The research reported here builds on previous studies, including our own previous experiments (Wierzba et al. 2023). We would thus like to clarify the relation to the present study and to stake out the scope and limitations of this paper.

The focus of our previous experiments was on providing experimental data concerning the effect of compositionality on syntactic flexibility of idioms in German and English. The methodological focus was on finding a way to estimate syntactic flexibility reliably. The theoretical focus was on assessing previously proposed hierarchies of syntactic structures with respect to their compatibility with idioms, like the hierarchy proposed for English by Fraser (1970) mentioned in Section 1.2 (a more detailed review of the structures and the motivation for selecting them can be found in Wierzba et al. 2023).

In the new experiments presented here, we reassess the compositionality factor, both methodologically and theoretically. Our first methodological goal is to reassess previous findings using an adjusted compositionality task. We also test a new set of materials that is larger and contains a wider range of idioms in order to test whether the compositionality effect is generalizable.

Our conceptual goal is to reconsider why and how we should expect compositionality to affect syntactic flexibility and what its explanatory limitations are. In the previous experiments, non-idioms just served as a baseline against which the idiom behavior was interpreted. In the present paper, we go a step further and ask whether we can find common sources of variability (beyond compositionality) in syntactic flexibility for both idioms and non-idioms. In the next section, we will discuss some recent research concerning parallels between idioms and non-idioms in various grammatical domains.

2.2.4 Parallels between idioms and non-idioms

Tabossi et al. (2009) report a piece of evidence revealing parallels between idioms and non-idioms. They found that providing a suitable context significantly raises the ratings for syntactically transformed idioms – except when the transformation violates a general formal requirement: in Italian, (non-idiomatic) bare nouns cannot appear in preverbal subject position. The same restriction was also found for bare nouns contained in idioms, and it was not alleviated by providing a context.

Horvath & Siloni (2009, 2019) observed, based on corpus searches of Hebrew and English idioms, that for most diatheses (voices) of the verb, there are idioms that uniquely occur in this diathesis (e.g., some idioms only occur in the adjectival passive: *caught in the middle* ‘fig. be between two opposing sides’, #catch X in the middle), but there are also diatheses for which this is not the case (e.g., verbal passive). Horvath & Siloni argue that idiomatic VPs are not stored as independent entries in the lexicon, but by “subentry storage”; for example, *kick the bucket* is stored as a subentry of the head *kick*. The argument is that if the idioms were stored as independent, holistic entries, it would be difficult to capture the systematic dependence on the head’s diathesis. Horvath & Siloni show that their observation also holds for non-compositional

idioms and thus, the representation of all idioms needs to have an internal structure to some extent (similar to non-idiomatic VPs). This proposal goes against Nunberg et al.'s (1993) proposal of distinguishing between compositional idioms with internal structure, and non-compositional idioms with holistic VP representations.

Bargmann & Sailer (2018) also argue against the view that non-compositional VP idioms should be represented as holistic chunks, based on web examples from German and English. They show that even non-compositional idioms are syntactically flexible to some extent (a finding corroborated by Fellbaum 2019), and that their syntactic modification underlies similar restrictions as in the case of compositional idioms and non-idioms. They propose an analysis in which even non-compositional idioms consist of individual lexical entries, but they differ from compositional idioms in that the entries involve semantic redundancy: for example, in their analysis of *kick the bucket,* the semantic contribution of *the* and *bucket* is contained in the contribution of kick.

Gehrke & McNally (2019) show that idioms that canonically contain a certain determiner can also occur with other determiners under certain circumstances; determiner variability is found with respect to definiteness and number, e.g. (based on examples by Bruening et al. 2018 and Everaert 2017):

(6) a. canonical: *to smell a rat*, fig. 'to sense something suspicious'
b. non-canonical (but attested): "Do we all smell many rats connected with this legislation?"

(7) a. canonical: *to kick the bucket*, fig. 'to die'
b. non-canonical: "Far more people pass on, kick *buckets*, […] than simply die."

Gehrke & McNally model such cases by separating the descriptive content (which can be compositional or non-compositional) from the potential to introduce discourse referents (anaphoric potential). The anaphoric potential interacts with compositionality: if an idiom is compositional for a speaker and it contains a DP that has its own figurative meaning, e.g., *a rat* in (6), the DP can be assumed to have its normal function of introducing a referent, in this case, a suspicious entity. When changed to 'many rats', several individuals are introduced as referents. As for an idiom like *kick the bucket,* whose descriptive content is non-compositional for most speakers, Gehrke & McNally (2019:794–796) discuss the possibility that it only introduces a referent at the level of events, but no referents at the level of individuals. The determiner could then express whether one or several events – as in (7) – are introduced.

The approaches above have in common that they stress commonalities between idioms and non-idioms and argue in favor of an analysis that reflects the parallels. Horvath & Siloni (2009; 2019), Bargmann & Sailer (2018), and Gehrke & McNally (2019) all propose models in which even non-compositional idioms are represented in a way that allows access to its individual

parts, for purposes of modifying diathesis, word order, or determiners. The latter two approaches nevertheless provide ways to represent differences between compositional idioms and non-compositional idioms without assuming that non-compositional idioms are unanalyzable chunks.

In this paper, we contribute experimental data to this line of research by studying parallels between idioms and non-idioms, in particular in Experiment 3.

3. Revisiting compositionality methodologically

Experiments 1 (on German) and 2 (on English) are replications of previous studies serving a two-fold purpose: we want to make sure that our previous findings (an effect of compositionality on the syntactic flexibility of idioms in German, but not in English) are robust. At the same time, we try out a new method: we collect compositionality ratings using a task that does not involve paraphrases. For this purpose, we consider it helpful to not change anything else about the design and materials to make sure that deviations (if any) can be attributed to the difference in methods.

The main hypothesis is that if compositionality determines the syntactic flexibility of idioms, then speakers' compositionality ratings for an idiom should be a good predictor for this idiom's syntactic flexibility, i.e., its acceptability in syntactically marked structures.

3.1 Experiment 1 (German)

3.1.1 Design and materials

The materials were adapted from our previous experiments on German (Wierzba et al. 2023). They included twelve idiomatic verb phrases, e.g., *das Handtuch werfen* (lit. 'to throw the towel', fig. 'to give up') and six non-idiomatic ones, e.g., d*en Bus verpassen* 'to miss the bus'. All idioms (throughout our Experiments 1–3) consist of a verb and a direct object. For a complete item list, see Appendix A.

All items were presented in the context shown in (8). In our previous experiments, we found that this type of context, which induces polarity focus in the target sentence, leads to higher ratings across various syntactic constructions than a broad-focus context. This helps to avoid the potential problem of comparing conditions that are perceived as unacceptable for independent reasons.

(8) Maria und Peter haben doch immer gegen die ungerechte Behandlung der Auszubildenden gekämpft. Haben sie inzwischen aufgegeben? *'Mary and Peter always used to fight against the unfair treatment of the trainees. Have they given up?'*

Each item was constructed in eight conditions (levels of the factor STRUCTURE):

(9) a. canonical word order:
Nein, die beiden würden bestimmt nie das Handtuch werfen!
no the two would definitely never the towel throw
'No, the two of them would definitely never throw in the towel!'

b. anaphor (pronominalization):
Nein, obwohl alle dachten, dass die beiden das Handtuch werfen würden, haben sie es doch nicht geworfen!
'No, even though everyone thought that the two of them would throw in the towel, they did not throw it in!'

c. prefield (fronting to the left periphery / topicalization):
Nein, das Handtuch würden die beiden bestimmt nie werfen!
no the towel would the two definitely never throw

d. left dislocation (LD):
Nein, das Handtuch, das würden die beiden bestimmt nie werfen!
no the towel PRON would the two definitely never throw

e. scrambling (over an adverbial):
Nein, die beiden würden das Handtuch bestimmt nie werfen!
no the two would the towel definitely never throw

f. passive:
Nein, so leicht wird das Handtuch nicht geworfen!
no so easily is the towel not thrown
'No, the towel is not thrown in so easily!'

g. nominalization:
Nein, den beiden ist bestimmt nicht zum Werfen des Handtuchs zu Mute.
no the two is definitely not to.the throwing of.the towel to spirit
'No, the two of them were definitely not in the mood for the throwing in of the towel.'

h. *which*-question: *'I heard that Mary and Peter have given up their fight.'*
Ach ja? Welches Handtuch sollen die beiden denn thrown have
oh yes which towel should the two PARTICLE geworfen haben?
'Oh yeah? And which towel are they supposed to have thrown in?'[3]

In our previous experiments, structures a/b/c/d/e were tested first, and a/c/d/f/g/h were tested in a follow-up experiment (some of the structures were included again to make sure that the results were comparable across experiments). Every participant saw each of the 18 items (12 idioms, 6 non-idioms) in all of the tested structures. The motivation was to reduce the risk that potential differences are due to individual variation with respect to how acceptable participants find each idiom in general rather than the syntactic manipulation.

[3] The *which*-question contains two modifications: fronting to the left periphery and insertion of *which*. As pointed out by a reviewer, each of the two modifications could influence the compatibility with idioms independently. We will not look at each of them separately here – to do that, it would be informative to include conditions with other types of determiner modification (e.g., *this towel*) for comparison.

In our replication, we combine the predecessor experiments into one study. To nevertheless stay close to the original designs, we created two lists: half of our participants saw each idiom and non-idiom in structures a/b/c/d/e, and the other half saw them in a/c/d/f/g/h. As a consequence, more judgments were collected for structures a/c/d, resulting in an unbalanced amount of observations per structure. In our Experiment 3 with new materials, we will use a different item distribution; here, our priority was to mirror the studies we replicate as closely as possible for comparability.

20 filler items were also adopted from the original materials. 10 contained a singular/plural manipulation of idioms (e.g., *in den sauren Apfel beiß*en, lit. 'to bite into the sour apple', fig. 'to do something necessary but unpleasant' / i*n die sauren Äpfel beißen* 'to bite into the sour apples'). The other 10 contained minimizers, which usually only occur under negation (*keinen Schimmer haben* 'have no clue', *#einen Schimmer haben* 'have a clue'), tested with and without negation. Each participant saw half of the fillers of each type in a condition which we expected to be acceptable (idioms containing a DP in its canonical singular/plural form or a minimizer licensed by negation) and half in a condition expected to be degraded (due to a number deviation or unlicensed minimizer).

In sum, each participant rated either 90 or 108 critical items (18 items in five/six conditions) and 20 fillers. The presentation order of stimuli was randomized.

3.1.2 Participants and procedure

The experiment was set up using the online questionnaire platforms SoSciSurvey (Leiner 2019) and L-Rex (Starschenko & Wierzba 2023). There were two parts to the experiment.

In the first part, participants provided compositionality ratings for 12 idioms. They were instructed that they would be asked questions about expressions with a figurative meaning. Participants were given two examples: *aus einer Mücke einen Elefanten machen* (lit. 'to turn a mosquito into an elephant', fig. 'to blow a small issue out of proportion') and *ins Gras beißen* (lit. 'to bite into the grass', fig. 'to die'). They were told that for the first example, it is possible to divide the expression up into two individual parts that each have their own figurative meaning, i.e., 'a mosquito' corresponds to a small problem, and 'an elephant' corresponds to a big fuss, and that this division is less straightforward for the second example. They then saw one idiom per page and answered the question 'Can this idiom be divided up into two parts that each have their own figurative meaning?'. There were four response options: (i) Yes, namely: (here, participants could enter paraphrases of the figurative meanings), (ii) Yes, but I cannot really express the individual meanings, (iii) No, (iv) I am not familiar with this idiom.

In the second part, participants were shown short dialogs including the same idioms and asked to rate the acceptability of the answer sentence on a scale from 1 (unacceptable) to 7 (acceptable).

There was a gap between the two parts, during which participants were asked to answer two unrelated questionnaires. These questionnaires served as a buffer between our two experiment parts and allowed us to make sure that the compositionality and acceptability tasks, which both involved the same idioms, did not directly follow each other. On average, the whole study took about one hour to complete.

48 native speakers of German took part. They were recruited via prolific.co and received £10 for participation. The pre-screening filters were set in such a way that participants spoke German as their first language, had been raised monolingually, and were born in and current residents of Germany.

3.1.3 Results

If a participant indicated that they were not familiar with an idiom in the pre-test, the participant's ratings for sentences containing this idiom were not included in the analysis of the acceptability ratings.

The compositionality ratings are shown in **Figure 2**. The acceptability ratings are illustrated in **Figure 3**.

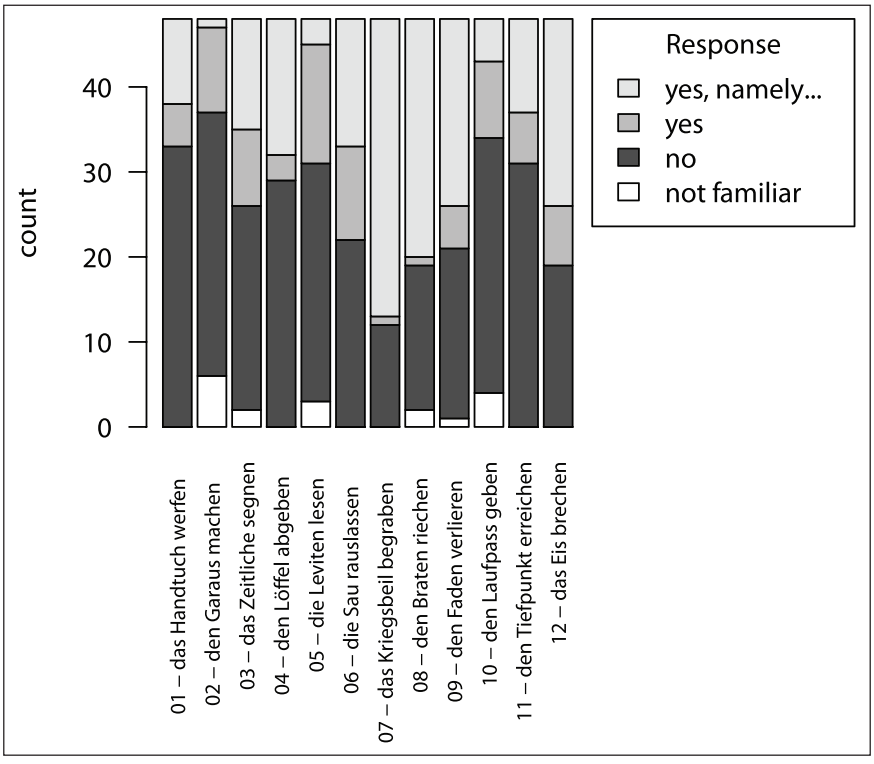

Figure 2: Compositionality judgments in Experiment 1: responses to the question 'Does each part of the idiom have its own individual figurative meaning?'

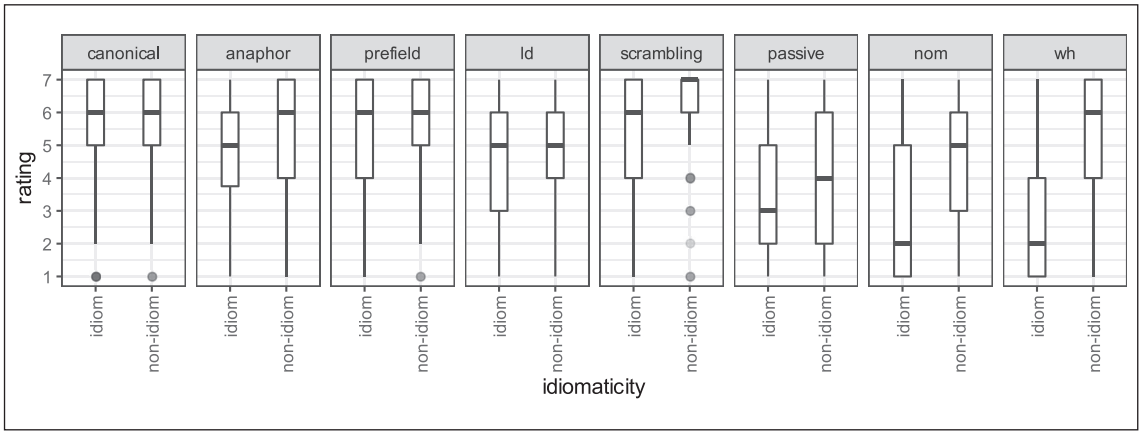

Figure 3: Boxplots representing the acceptability results of Experiment 1, split by idiomaticity and structure.[4]

For statistical analysis, the factor idiomaticity (idioms vs. non-idioms) was sum-coded in order to treat the two levels symmetrically. The factor structure was treatment-coded with canonical as the baseline to which the other structures are compared. This contrast-coding allows us to test whether there is an acceptability contrast between idioms and non-idioms in the canonical baseline, and whether this contrast is larger in the other structures.

We ran linear mixed models (LMMs) as well as cumulative link models (CLMs). For reasons of space, we only report the main LMM results in the paper. We provide the detailed model specifications and output of the LMM analyses in Appendix C. The CLM results are available in our Open Science Framework (OSF) repository under https://doi.org/10.17605/OSF.IO/JPUFT. It has been shown that linear mixed models can increase Type I and Type II errors and distort estimates of effect size when applied to ordinal rating data (Liddell & Kruschke 2018; Veríssimo 2021). The theoretical conclusions that we draw in this paper are based on tests that consistently yielded significant results in both the LMM and CLM analysis. We summarize and discuss any deviations that occurred between the model types in Appendix B.2.

According to an LMM[5] fit to the whole data set, no simple effect of idiomaticity was found, i.e., no significant idiom/non-idiom contrast in the canonical baseline. As for the interaction between idiomaticity and structure, a larger contrast between idiom and non-idiom than in the canonical baseline was found for anaphor ($t = -2.54$, $p = 0.02$), prefield ($t = -3.65$, $p = 0.002$), LD ($t = -3.14$, $p = 0.005$), nominalization ($t = -4.39$, $p < 0.001$), and *which*-question ($t = -5.23$,

[4] The horizontal line represents the median. The box represents the interquartile range (IQR), i.e., the distance between the upper and lower quartiles. The whiskers represent the range of data points falling within a distance of 1.5 times the IQR above/below the upper/lower quartiles. Circles represent data points outside of this range.

[5] We followed the recommendations for identifying parsimonious models by Bates et al. (2015a). We used the R packages lme4 and lmerTest (R Core Team 2016; Bates et al. 2015b; Kuznetsova et al. 2017).

$p < 0.001$), while the contrast was not significantly different from the canonical baseline for scrambling and passive.

An additional LMM was fit to the data subset including only idioms in order to test the influence of compositionality on flexibility. The compositionality measure that we used as a dependent variable was the proportion of positive responses to the question that we asked in the compositionality task (excluding "I am not familiar with this idiom" responses). For example, the compositionality value for *das Handtuch werfen* 'to throw in the towel' was 0.32 (32% of participants said that each part of the idiom had its own figurative meaning), while it was 0.75 for *das Kriegbeil begraben* 'to bury the hatchet'. We included this compositionality rating as a linear predictor in our model. No simple effect of compositionality was found, i.e., no significant linear effect of compositionality on the ratings in the canonical baseline. As for the interaction between compositionality and structure, a larger effect of compositionality than in the canonical baseline was found for anaphor ($t = 13.78$, $p < 0.001$), prefield ($t = 2.45$, $p = 0.004$), LD ($t = 2.30$, $p = 0.04$), scrambling ($t = 3.04$, $p = 0.01$), and *which*-question ($t = 2.45$, $p = 0.03$), while the effect was not significantly different from the canonical baseline for passive and nominalization.

The estimate of the effect of compositionality is illustrated in **Figure 4**. The crucial interaction between compositionality and structure is illustrated by the steepness of the fitted lines and the width of the confidence bands. For example, the steeply increasing fitted line in the scrambling structure with a narrow band indicates that idioms with a higher compositionality value tend to correspond to higher acceptability of scrambling with relatively high confidence. In contrast, the broader confidence band and the flatness of the line in the passive structure indicates that there is little evidence for a linear relation between compositionality and the acceptability of passivization in our data – there are some highly compositional idioms that can easily be passivized, but also some that cannot be passivized felicitously, and the same holds for highly non-compositional idioms.

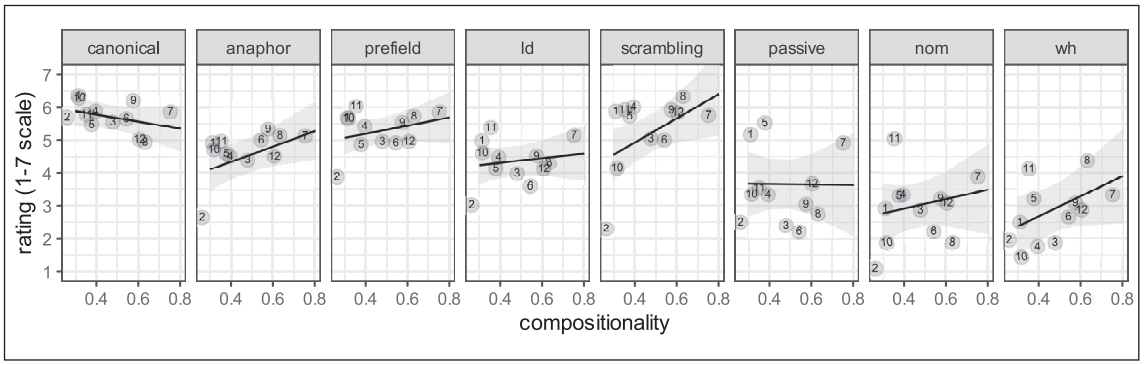

Figure 4: Estimate of the effect of compositionality (as a linear predictor) on the acceptability of the tested structures in Experiment 1. The circles additionally show the observed by-item means for each structure.

All significant effects reported above according to the LMM were also significant in the CLM analysis. We will only interpret these effects in the discussion below. We summarize and discuss any additional significant effects found in the CLM but not in the LMM in Appendix B.2.

3.1.4 Discussion

Our results replicate several of Wierzba et al.'s (2023) findings, with the crucial difference that this time, compositionality was measured empirically by collecting judgments from participants.

First, there is a significant gap between the acceptability of idioms and non-idioms in most of the tested syntactically marked structures. This supports the view that on average, there is an empirical difference in the syntactic behavior between idioms and non-idioms. Second, our results are in line with our previous finding that compositionality of German idioms is a significant predictor of their syntactic flexibility; we have replicated this using an empirically grounded measure of compositionality. Our results furthermore show that it is worth having a differentiated look at syntactic flexibility: idioms with a higher compositionality rating receive higher acceptability ratings in marked structures on average, but a by-structure look reveals that this does not necessarily hold for every tested structure in the same way. A significantly larger effect of compositionality (in comparison to canonical word order) was found consistently (in both the LMM and CLM analyses) for the following constructions: anaphor, prefield, LD, scrambling, and *which*-question.

Our results substantiate the view that there is not only variability in the syntactic flexibility of idioms, but also among non-idioms. For most of the tested structures, the interquartile ranges of the non-idioms suggest a similarly broad distribution of data points as for idioms. An approach in terms of compositionality can only explain the variability in the behavior of idioms, as non-idioms do not vary in compositionality. The question thus remains whether there are factors that systematically determine the syntactic flexibility of the non-idioms, and whether it is the right approach to assume that they are different from the factors determining the flexibility of idioms. We will pursue this question further in Experiment 3.

An additional post-hoc observation is that a subset of the tested structures seems to correlate with each other with respect to the acceptability of individual idioms. Visual inspection of **Figure 4** suggests that the more acceptable an idiom is with prefield fronting, the better it works also with LD and scrambling. For initial notes and hypotheses on the observed correlations, see Appendix B.1 and our data repository; the discussion of these correlations is beyond the scope of the present paper, but is worth further inspection in future work.

A closer look at the compositionality ratings in comparison to the introspective categorization used in our previous experiments reveals some informative deviations. While most of the items previously categorized as non-compositional (items 1–6) received more 'no' responses than those previously categorized as compositional (7–12), some of the latter fall outside of the pattern, a.o.

den Tiefpunkt erreichen (lit. 'to reach the lowest point', fig. 'to be in the worst possible situation'). A special property of this item is that *Tiefpunkt* can be used in a metaphorical sense independently of the verb *erreichen*; thus, as pointed out by a reviewer, it is only partially idiomatic. A limitation of our method is that we ask participants whether both parts of the idioms (verb/object) have their own *figurative* meaning. In cases like *den Tiefpunkt erreichen*, we get a high proportion of 'no' responses; however, the reason is not that the verb and the object do not combine compositionally (they definitely do), but partial idiomaticity. It is thus probable that our method underestimates compositionality in cases like this, and partially idiomatic expressions should be avoided.[6]

Another note about the results in comparison to the previous experiments is that the average acceptability ratings are relatively similar numerically.[7] This alleviates the potential worry (raised by two anonymous reviewers) that the compositionality rating task might have influenced participants' acceptability ratings.

3.2 Experiment 2 (English)

3.2.1 Design and materials

The materials were adapted from our previous experiment on English (Wierzba et al. 2023); the items were based on lists of compositional/non-compositional idioms provided by Gibbs & Nayak (1989). They included twelve idiomatic verb phrases, e.g., *pop the question* (fig. 'to propose') and six non-idiomatic ones, e.g., *forget the timer*. For a complete item list, see Appendix A. Each item was constructed in six conditions:

(10) Meghan is really excited. Do you think Harry asked her to marry him?
   a. canonical word order: Of course not, he would definitely never pop the question!
   b. anaphor: I'd say so… even though no one thought he would ever pop the question, he obviously did pop it.
   c. nominalization with "of": Yes, he did, but I don't really want to talk about Harry's popping of the question at the moment.
   d. nominalization without "of": Yes, he did, but I don't really want to talk about Harry's popping the question at the moment.
   e. passive: Of course not, the question would definitely never be popped by such an incorrigible player!
   f. cleft-like: Of course not, the question is something that he would definitely never pop!

[6] We reran the LMMs without the item and provide the results in the OSF repository. The only difference to the complete dataset was that the interaction between compositionality and structure reached significance for nominalization when excluding the item *den Tiefpunkt erreichen*.

[7] For comparison, when we pool all data from the three previous German experiments, the medians by condition are: canonical: 6 for idioms / 7 for non-idioms, prefield: 5/6, LD: 4/5, scrambling: 5/6, anaphor 5/6, passive: 3/5, nominalization 2/3.5, *which*-question: 2/5.

As in the original study we are replicating, each participants saw each of the 18 items (12 idioms and 6 non-idioms) in all of the tested structures a–f.

20 filler stimuli were adopted from the original study, which were designed in the same way as the German fillers in Experiment 1: 10 contained a singular/plural manipulation (e.g., *turn over a new leaf / turn over new leaves*), and 10 contained minimizers with/without negation (e.g., *I did not get a wink of sleep / #I got a wink of sleep*).

In sum, each participants rated 108 critical items (18 items in six conditions) and 20 fillers. The presentation order was randomized.

3.2.2 Participants and procedure

The experiment was set up in a similar manner as described for Experiment 1. There was a gap of at least two hours between the compositionality and acceptability rating tasks. In contrast to Experiment 1, participants were not asked to complete other questionnaires during the gap. In sum, the two study parts took about 40 minutes to complete. The pre-screening filters that we set on prolific.co only allowed speakers who spoke English as their first language and were current residents of the UK to take part. Participants received £6.25 for participation.

3.2.3 Results

If a participant indicated that they were not familiar with an idiom in the pre-test, their ratings for sentences containing this idiom were not included in the analysis of the acceptability ratings.

The compositionality ratings are shown in **Figure 5**. The acceptability ratings are illustrated in **Figure 6**.

According to a LMM fit to the whole data set, there was a simple effect of idiomaticity. As for the interaction between idiomaticity and structure, a larger contrast between idiom and non-idiom than in the canonical baseline was found for passive ($t = -2.52$, $p = 0.02$) and the cleft-like structure ($t = -5.00$, $p < 0.001$). The contrast was not significantly different from the canonical baseline for anaphor and nominalization with "of". A significant deviance in the opposite direction (smaller contrast between idiom and non-idiom than in the canonical structure) was found for nominalization without "of" ($t = 2.19$, $p = 0.04$).[8]

An additional LMM was fit to the data subset including only idioms in order to test the influence of the factor compositionality. The same compositionality measure was used as in Experiment 1. A simple effect of compositionality was found: ratings were higher for more compositional idioms in the baseline. This means that idioms that were perceived as more compositional tended to be judged as more acceptable even in sentences with canonical word order, perhaps due to

[8] This was the only significant effect in the LMM analysis which was not significant in the CLM analysis.

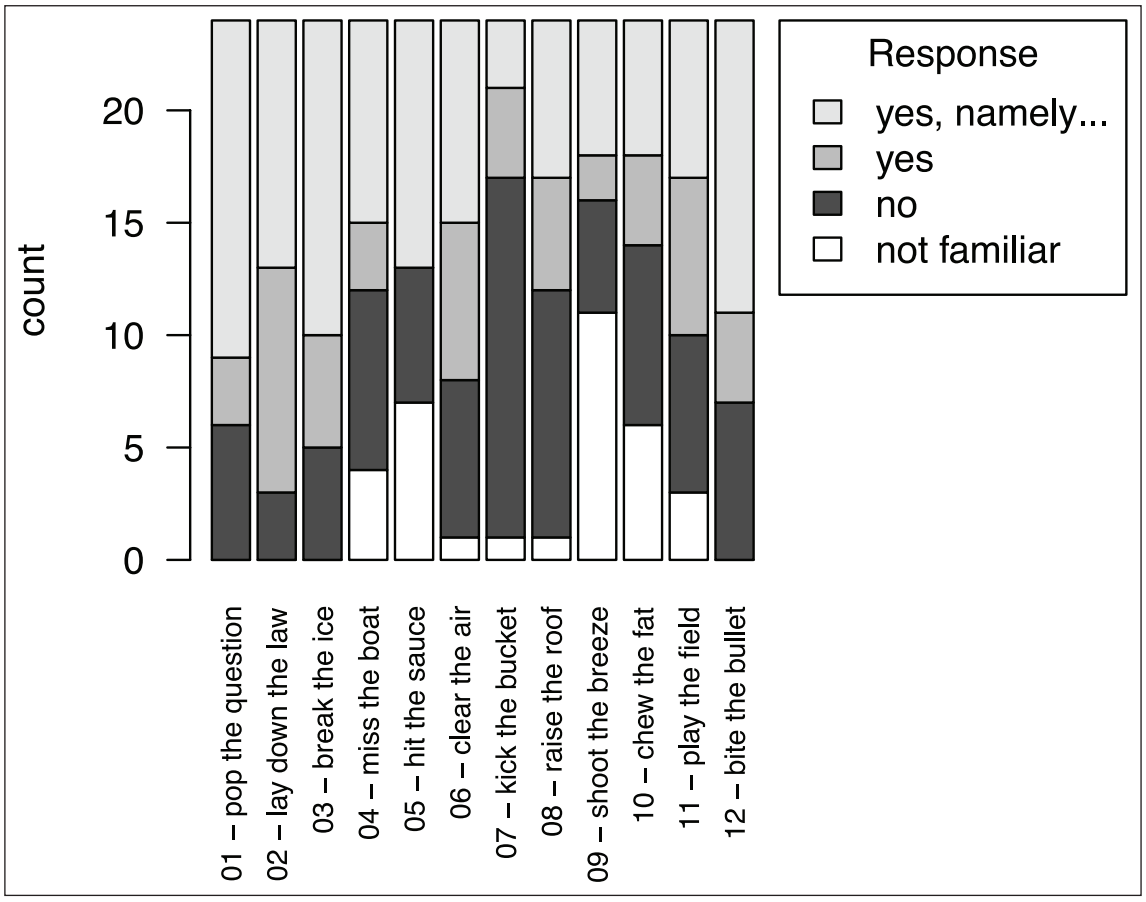

Figure 5: Compositionality judgments in Experiment 2: responses to the question 'Does each part of the idiom have its own individual figurative meaning?'

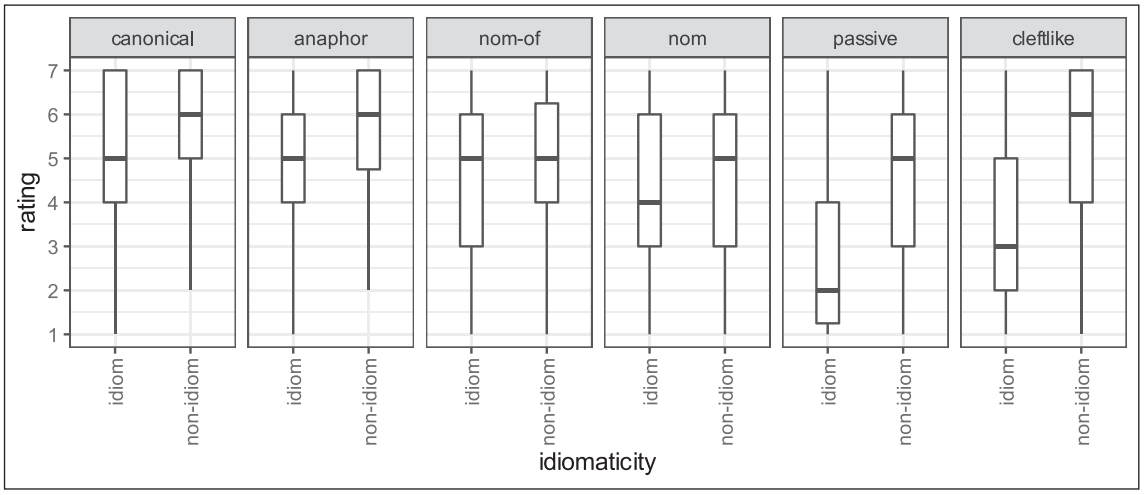

Figure 6: Boxplots representing the acceptability results of Experiment 2, split by idiomaticity and structure.

higher familiarity (we will come back to this below). As for the interaction of compositionality with structure, a larger positive effect of compositionality than in the canonical baseline was not found for any of the tested structures. The effect of compositionality did not differ from the one found in the canonical baseline for anaphor, passive, and the cleft-like structure. A significant deviance from the canonical baseline (an in comparison less positive effect of compositionality) was found for nominalization with "of" ($t = -2.87$, $p = 0.004$) and nominalization without "of" ($t = -4.15$, $p < 0.001$). The estimate of the effect of compositionality is illustrated in **Figure 7**. As in Experiment 1, a steep fitted line with a narrow confidence band indicates a strong linear relation between compositionality and the acceptability of the syntactic structure at hand. However, it is important to note that in contrast to Experiment 1, there already is a linear relation, i.e., a steep fitted line in the canonical baseline condition. Thus, the steepness of the fitted line in the other conditions, e.g., the passive structure, needs to be compared to the canonical baseline: only if the line was significantly steeper for passive than for canonical, it would indicate that higher compositionality facilitates passivization; this is however not the case here, neither for passive nor for any of the other marked structures; in some of the structures (the nominalization structures) the positive linear relation is even significantly less pronounced.

The same pattern of significant and non-significant effects as in the LMMs was found in the CLM analysis with the exception of the deviance noted in footnote 9, which is not relevant for the discussion below.

3.2.4 Discussion

Our results confirm previous findings of Wierzba et al. (2023), but also go against some previous claims. In particular we failed to find the positive effect of compositionality on syntactic flexibility that Gibbs & Nayak (1989) reported for English.

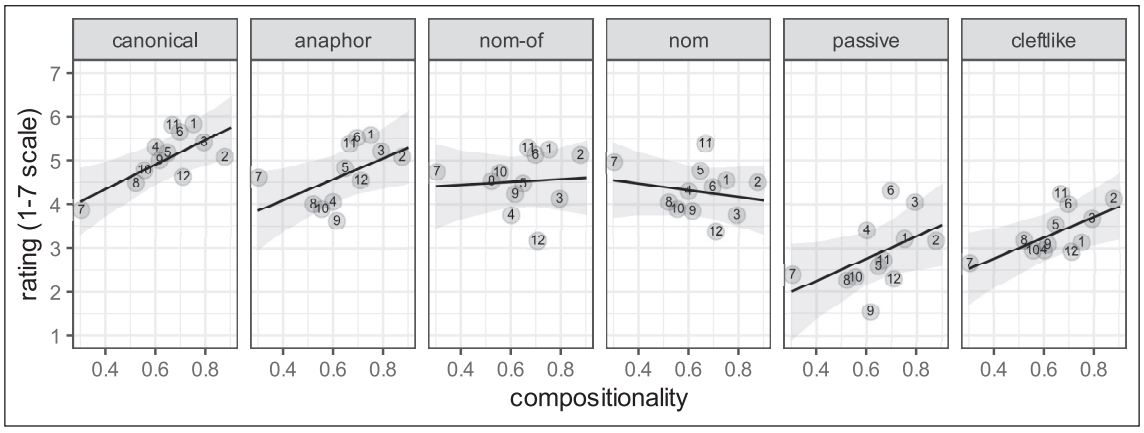

Figure 7: Estimate of the effect of compositionality (as a linear predictor) on the acceptability of the tested structures in Experiment 2. The circles additionally show the observed by-item means for each structure.

As in Experiment 1, we do see a gap between the acceptability of idioms and non-idioms in a part of the tested syntactically marked structures (passive, cleft-like) that is larger than in the canonical structure. This lends further support to the view that there is indeed an empirical difference in the syntactic flexibility of idioms and non-idioms that requires an explanation; however, as in Experiment 1 on German, a differentiated look is useful, as the gap is not present (at least not in a stronger form than in the canonical baseline) in all syntactically marked structures.

Like Wierzba et al. (2023), we found a general effect of compositionality for English in our Experiment 2: sentences with idioms judged as more compositional tended to be more acceptable. However, again, this effect was found to already be present in the baseline structure with canonical word order, and it was similar in size and partially even smaller in all of the syntactically marked structures that we tested. Thus, this experiment does not provide evidence that compositionality is the factor that explains the idiom/non-idiom gap in syntactic flexibility. The results confirm that it is important to interpret any contrasts found in the syntactically marked constructions against a canonical baseline (which was lacking in Gibbs & Nayak's 1989 study) – otherwise, contrasts between various idioms that arise due to independent factors and have nothing to do with syntactic flexibility might be overinterpreted.

As for the question of why we see a compositionality effect even in the canonical baseline, we speculate that familiarity might play a role. Even though we excluded participants' data points for idioms that they judged as non-familiar, there still might be gradient differences within the remaining idioms, potentially leading to lower acceptability ratings as a result of feeling unsure about some of the stimuli. If the less compositional idioms happened to be less familiar to the participants, this could be the reason for the observed correlation between compositionality and acceptability.

It can further be noted that the compositionality judgments participants gave in this experiment do not fully line up with the categorization reported by Gibbs & Nayak (1989) and adopted in Wierzba et al. (2023). We provide a more detailed comparison in Appendix B.3.

In 3.1, we discussed the caveat that one of our German items was only partially idiomatic. In the German example *den Tiefpunkt erreichen* (lit. 'to reach the lowest point', fig. 'to be in the worst possible situation'), *Tiefpunkt* can signify a bad situation independently of *erreichen*. The English items of Experiment 2 illustrate a similar type of item for which compositionality might have been underestimated (we thank a reviewer for underscoring this point). The idiom *miss the boat* can be paraphrased as *miss an opportunity*. *Boat* can signify *opportunity* in other examples such as *The boat has sailed* or *Look, the boat's gone, let's just cut our losses and go*. For this reason, participants might have hesitated to say that *miss* has its own figurative meaning in *miss the boat,* as *miss* can be interpreted literally here if *boat* corresponds to *opportunity*. Similar considerations apply to the verb *clear* in *clear the air* and the nouns *question* and *law* in *pop the question* and *lay down the law*. However, visual inspection of the compositionality judgments in **Figure 5** shows a

much lower proportion of negative responses (<50%) to the question whether both idiom parts have a figurative meaning for all of these English examples than for *den Tiefpunkt erreichen* in Experiment 1. This suggests that the availability of a literal interpretation of the DP/verb did not necessarily prompt participants to give a negative response in the compositionality task, but it is nevertheless a point that is worth being considered in future work.

Finally, we do not want to argue that the results of Experiment 2 rule out the possibility that compositionality affects the syntactic flexibility of idioms in English – there are various potential reasons for the absence of an effect: it could be due to methodological limitations like the sample size or, e.g., the fact that we only tested for a linear relation between the two properties. It could also be due to differences between the tested constructions. Both our German and English materials included several structures that we assumed to potentially depend on compositionality for semantic reasons: for English, these included anaphor, passive (argued to be linked to topicality/referentiality a.o. by Nunberg et al. 1994 and reported to depend on compositionality by Gibbs & Nayak 1989), and the cleft-like construction (linked to focus/exhaustivity; see Wierzba et al. 2023 for more detailed discussion). However, a part of the German structures for which we found a compositionality effect did not have a direct counterpart in the English materials (a.o., prefield, LD, scrambling) – if manipulations of linear order are the type of structure for which one is most likely to find a compositionality effect in our type of experimental design, there was less opportunity to find it in the English experiment.

Despite these limitations, we would like to stress the insight that the choice of the task used to measure compositionality and syntactic flexibility influences the results of the task, as evidenced by the divergent conclusions of Gibbs & Nayak 1989 vs. Wierzba et al. (2023) and the present study. Thus, previous findings of a positive effect of compositionality on flexibility should be taken with a grain of salt, and the question seen as re-opened.

3.3 Interim summary of Experiments 1–2

In Section 1, we formulated the following research questions: (Qi) Why are some idioms more flexible than others? (Qii) Why are some non-idioms more flexible than others? (Qiii) Is there a gap in syntactic flexibility between idioms and non-idioms?

With respect to (Qiii), we have seen that idioms are significantly less acceptable in (some) marked syntactic constructions than non-idioms in both Experiment 1 and 2, i.e., there is indeed a contrast in syntactic flexibility between the two types of expressions.

One explanation would be that idioms are less flexible because their meaning is less compositional on average. If this is correct, we would expect to find a connection between compositionality and syntactic flexibility within the idioms, which would also constitute an answer to question (Qi). Our findings for German (Experiment 1) support this view, in contrast to the ones for English (Experiment 2).

In the next section, we will follow up on the German results. One of the goals is to take a first step towards addressing question (Qii): compositionality has the potential to explain the variability in syntactic flexibility of idioms, but it is not straightforwardly extendable to non-idioms (which we take to be compositional by definition). We will thus take a look at further factors which have the potential to explain (some of) the variability in the syntactic flexibility of both idioms and non-idioms.

4. Extending the data set and revisiting compositionality conceptually

The main goals of Experiment 3 are to test whether our findings so far are generalizable, and take a step towards investigating parallels between non-idioms and idioms.

In Experiments 1–2, all items included a definite DP. In Experiment 3, we extend the empirical domain by adding further DP/noun types, including indefinite and incorporated nouns. The design with a higher number and greater variety of expressions is motivated by the aim to investigate factors that might influence both idioms and non-idioms, focusing on definiteness and referentiality.

If the compositionality effect found in Experiment 1 is generalizable, we should find it again in this new set of items. Our assumptions and expectations about definiteness and referentiality will be discussed in 4.1.

4.1 Definiteness and referentiality

4.1.1 Referentiality in the sense of anaphoric potential

In Experiment 3, we test indefinite and definite DPs, which are related to reference in several ways. First, they can introduce discourse referents (Karttunen 1969), which can then be picked up by pronouns (e.g., *Mary ate an apple. It tasted great.*); we will refer to this ability as 'anaphoric potential'. Definite expressions presuppose a (unique or familiar) referent to already be present in the discourse and point to it (Heim 1991; Schwarz 2009). However, in appropriate contexts, it is often possible for speakers to accommodate a unique referent even if it has not been introduced explicitly (Singh et al. 2016); in this way, definite expressions can also introduce referents. Thus, we assume that in principle, both definite and indefinite DPs have anaphoric potential.

What about idiomatic DPs? As argued by Gehrke & McNally (2019), definite and indefinite DPs that are contained in idioms can in principle also introduce referents at the level of individuals (i.e., they have anaphoric potential), as long as the idioms are compositional (see Section 2.2.4). Non-compositional idioms, on the other hand, are more limited in this respect. Since pronominalization is one of the phenomena that has been used to estimate syntactic flexibility, anaphoric potential is relevant for our research questions.

4.1.2 Referentiality in the sense of specificity

A further DP property related to reference is specificity. *Specificity* is a property that is often discussed in connection with different readings of indefinite DPs, as in (11).

(11) a. A student cheated on the exam. His name is John. *specific*
b. A student cheated on the exam. We are all trying to figure out who it was. *non-specific*

According to Fodor & Sag (1982), the difference between the readings is whether the speaker has a certain referent "in mind". Under this view, definite DPs are often specific, because they usually refer to an already established discourse referent.[9] For indefinite DPs, we assume that a specific reading can be enforced by the context as in (14b), but that it is not always a salient interpretation.

And idiomatic DPs? Analogously to the discussion of compositionality and anaphoric potential above, we also think that compositionality is a prerequisite for interpreting a part of an idiom as specific: only if an expression has an individual (figurative) meaning of its own within the idiom, speakers can have a certain referent for it in mind.

Specificity is relevant for the discussion of syntactic flexibility of idioms, because it has been argued to have an impact on word order modifications, in particular scrambling (Diesing 1990; Lenerz 2001; Frey 2004; Fanselow 2018). Frey (2004) gives the example in (12) to illustrate that scrambling the object over the adverbial results in a specific interpretation of *ein Kind* 'a child', which is infelicitous in the provided context that enforces an unspecific interpretation.

(12) Context: *Hans and Maria got married.*
#Ich denke, dass ein Kind bald kommen wird.
I think that a child soon come will

There are approaches according to which there is a relation between scrambling and other structures, which might therefore also be impacted by specificity: e.g., it has been proposed that prefield fronting of objects involves a step of object scrambling (Müller 2004), and that LD, in turn, involves prefield fronting (Grohmann 2000).

4.1.3 The relation between compositionality, referentiality, and syntactic flexibility: assumptions and expectations

We will use the term *referentiality* in a dual sense here: We take an expression to be referential if it has anaphoric potential or it is specific. In this sense, referentiality is a semantic property that idioms and non-idioms possess to varying degrees. Thus, it has the potential to provide a unified explanation for (a part of) the variability found in the syntactic flexibility of idioms as well as non-

[9] See Kratzer (1998) for a discussion of the relation between specificity and other semantic notions like scope, and von Heusinger (2002) for discussion of definiteness and specificity as in principle orthogonal categories.

idioms. Under this view, we would expect compositionality to correlate roughly with syntactic flexibility – however, not due to a direct relationship between these two properties, but due to a more complex dependency: compositionality of an idiom is a prerequisite for referentiality, and referentiality, in turn, is a prerequisite for some syntactic modifications.

If the reasoning in 4.1.1–4.1.2 is correct, we would expect similar behavior of definite/indefinite DPs with respect to pronominalization. As for scrambling, a contrast between definite/indefinite DPs should be detectable if a specific reading is not salient in the given context. For idioms, the availability of pronominalization and scrambling should additionally depend on compositionality.

4.1.4 Incorporated nouns: non-idioms with limited referentiality

Besides idioms, there are also nominal elements which have special referential properties even though they do not involve figurative meaning. An example of this type of expression that we will focus on in Experiment 3 are nouns in German predicates like *Zeitung lesen* 'read (a/the) newspaper'. We follow Frey (2015:258) in assuming that such predicates involve a "very close syntactic junction" between the verb and the noun and in referring to it as incorporation. This type of construction combines a verb with a (semantically/pragmatically) "typical" direct object. A special property is that the object occurs in a bare (articleless) form, even if it is a singular count noun – usually, only mass nouns and indefinite plurals can occur without an article in German. The construction shows several syntactic characteristics differing from typical complex transitive VPs, for example when it comes to the interaction with sentence negation (see Frey 2015 for details and further evidence).

Modarresi & Krifka (2021) report speaker judgments showing a gradient pattern of anaphoric potential for a set of items with a bare singular noun; e.g., while the noun in *Zeitung lesen* completely lacks anaphoric potential as illustrated in (13), the noun in other examples like *Kuchen backen* 'bake (a/the) cake' can serve as an antecedent to some (limited) extent.

(13) Peter hat heute *Zeitung* gelesen. **Sie* war sehr interessant.
Peter has today newspaper.F.SG read PRON.F.SG was very interesting
'Peter has read (a/the) newspaper today. It was very interesting.'

With respect to specificity, we follow Frey (2015:230) in assuming that incorporated nouns are non-specific.

Our motivation for including incorporated nouns in Experiment 3 is that they can potentially help to disentangle referentiality from idiomaticity. If the assumptions above are correct, verb phrases with incorporated nouns should show limited syntactic flexibility, even though they are non-idiomatic.

4.2 Experiment 3 (German)

4.2.1 Participants and procedure

The compositionality rating task was adjusted in comparison to Experiments 1–2. First, we asked which of the idioms participants were familiar with. Then, we only asked for compositionality ratings for the familiar idioms. We asked separately whether each of the two parts of the idiom (verb/DP) had its own figurative meaning. The motivation for this adjustment was that it highlights which idiom part participants failed to assign an individual meaning to. To nevertheless make the results comparable across experiments, we will assume that two positive responses in Experiment 3 ("Does the verb/DP have its own meaning?") correspond to a positive response in Experiments 1–2 ("Does each part have its own figurative meaning?") in our analysis. At least one negative response in Experiment 3 will correspond to a negative response to the single question in Experiment 1–2. We will return to the additional information provided by the more fine-grained task in the discussion below.

The acceptability rating task was the same as in Experiments 1–2.

60 native speakers of German (again recruited via prolific.co) took part. They received £9 for participation.

4.2.2 Design and materials

16 new idiomatic VPs were used. Eight of them included a definite DP, e.g., *die Wogen glätten* (lit. 'to smooth the waves', fig. 'to calm things down'), and eight included an indefinite DP, e.g., *eine dicke Lippe riskieren* (lit. 'to risk a fat lip', fig. 'to say something cheeky'). In addition, 16 non-idiomatic VPs were used. Four included a definite DP, e.g., *die Preise senken* ('to lower the prices'), four included an indefinite DP, e.g., *eine Mahlzeit zubereiten* ('to prepare a meal'), and eight included a non-figurative incorporated noun, e.g., *Zeitung lesen* ('to read (a) newspaper').[10] We wanted to balance the number of expressions that we assume to have / not have anaphor potential (indefinite and definite DPs vs. incorporated nouns). In our items with an indefinite DP, a specific reading of the DP was not salient in the provided contexts; e.g., in '[…] Do you think that Leon will send us greetings via mail? – No, he would definitely never send us a postcard', it is not a salient interpretation that the speaker has a certain postcard in mind that Leon would never send. A complete item list is provided in Appendix A.

A different approach to item distribution was taken in Experiment 3. Two lists were created. Every questionnaire contained each item in four out of the eight structures. For each item, it was randomly determined which of the structures belonged to which list. This way, the same amount of data points was collected for each structure, resulting in a more balanced distribution than

[10] We will return to the distinction between indefinite DPs and incorporated nouns in Section 5.4.

in Experiment 1. The distribution scheme did however not guarantee that every participant saw the same number of items per condition. The presentation order of the stimuli was randomized.

As motivated in Section 4.1, the most important structures for evaluating the hypotheses on compositionality and referentiality are those that involve a word order modification potentially linked to referentiality (scrambling, anaphor). However, to make a direct comparison to the previous studies possible and to see which findings carry over to the new data set, as well as for exploratory purposes, we included all syntactic structures from Experiments 1–2.[11]

4.2.3 Results

A summary of the compositionality ratings is shown in **Figure 8**. The acceptability ratings are shown in **Figure 9**. If a participant indicated that they were not familiar with an idiom in the pre-test, their ratings for sentences containing this idiom were not included in the analysis of the acceptability ratings.

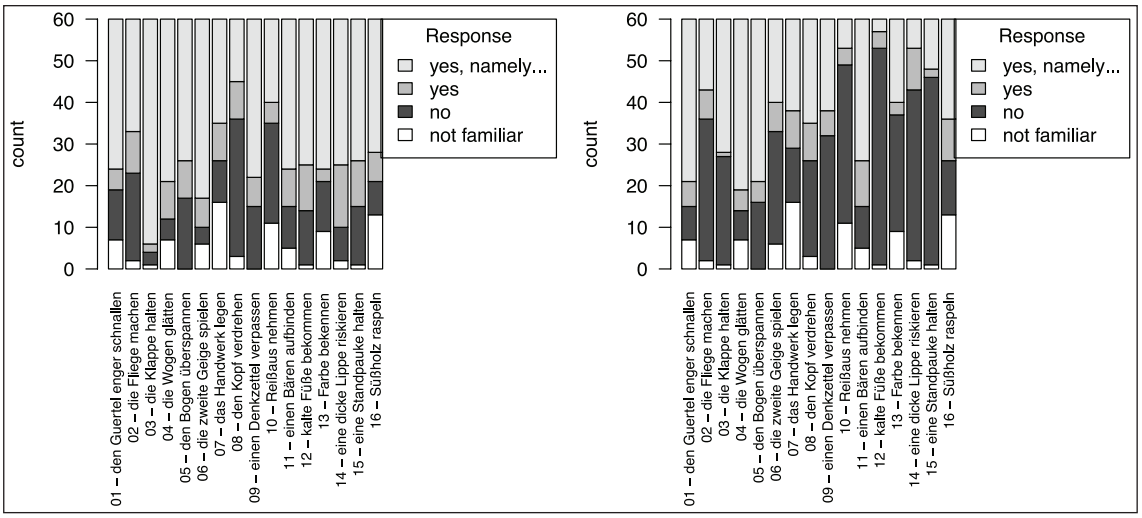

Figure 8: Compositionality judgments in Experiment 3: responses to the question whether the DP (upper plot) / the verb (lower plot) has its own individual figurative meaning.

The factor idiomaticity (idioms, non-idioms) was sum-coded and the factor structure was treatment-coded with canonical as the baseline. For the additional factor DP type (definite, indefinite, incorporated), forward difference coding was used, allowing us to compare definite DPs to indefinite ones and indefinite DPs to incorporated nouns.

[11] The passive condition was constructed differently from Experiment 1. We used sentences like *Wenn die Wogen von jemandem geglättet wurden, dann bestimmt nicht von Maria*, lit. 'If the waves were smoothed by somebody, then certainly not by Mary. 'The reason for this change was to provide more motivation for using the passive (as a way of focusing Mary); we felt that such a motivation was lacking in Experiment 1.

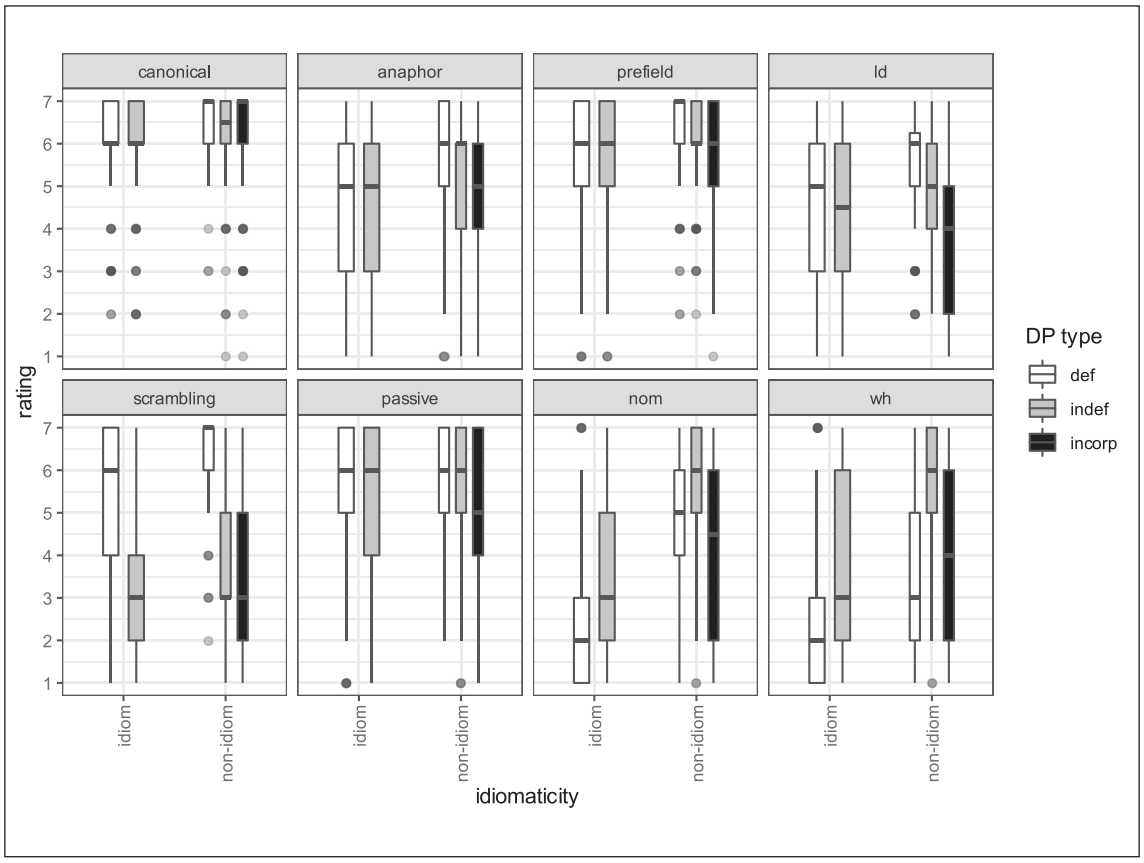

Figure 9: Boxplots representing the acceptability results of Experiment 3, split by idiomaticity, DP type, and structure.

We ran three analyses. First, we tested the effects of structure, idiomaticity, and DP type in the data subset of definite and indefinite DPs in order to see whether definiteness had a similar effect on idioms and non-idioms; incorporated nouns were excluded here, because they are a special case of non-idioms. Second, we tested the effect of compositionality within the idiom subset. Third, we tested whether there was a difference between the indefinite DPs and incorporated nouns within the non-idioms subset.

In the first analysis, we fit a model to the subset of the data with definite/indefinite DPs (excluding the items incorporated nouns), with structure, idiomaticity, and DP type as fixed factors. No simple effect of idiomaticity was found, i.e., no significant idiom/non-idiom contrast in the canonical baseline. No simple effect of DP type was found, either, i.e., no significant definite/indefinite contrast in the canonical baseline. There also was no significant interaction between idiomaticity and DP type in the canonical baseline. As for the interaction between idiomaticity and structure, a larger contrast between idiom and non-idiom than in the canonical baseline in the direction of higher ratings for non-idioms was found for anaphor ($t = -3.75$,

$p < 0.001$), prefield ($t = -2.42$, $p = 0.02$), LD ($t = -2.49$, $p = 0.017$), nominalization ($t = -5.62$, $p = 0.001$), and *which*-question ($t = -3.51$, $p = 0.001$), while the contrast was not significantly different from the canonical baseline for passive. As for the interaction between definiteness and structure, a larger contrast between definite and indefinite DPs than in the canonical baseline towards higher ratings for definite was found for scrambling ($t = 7.81$, $p < 0.001$). A significant effect in the opposite direction (towards higher ratings for indefinite) was found for nominalization ($t = -2.58$, $p = 0.02$) and *which*-question ($t = -5.20$, $p < 0.001$), while the contrast was not significantly different from the canonical baseline for the other structures. The latter finding was qualified by a three-way interaction between structure, idiomaticity, and DP type: in *which*-questions, the contrast between definite and indefinite DPs was larger in non-idioms than in idioms ($t = 2.92$, $p = 0.007$). No other three-way interactions were significant.

In the second analysis, we fit an additional model to the data subset including only idioms in order to test the influence of the factor compositionality and its interaction with structure and DP type. The compositionality measure that we used as a dependent variable was the proportion of cases in which a positive response to both questions asked in the compositionality task was given (concerning an individual figurative meaning of the DP and of the verb). No simple effect of compositionality was found, i.e., no significant linear effect of compositionality on the ratings in the canonical baseline (rather, there was a trend towards a negative effect in the canonical baseline). No significant effect of DP type was found, either, i.e., no significant contrast between definite and indefinite in the canonical baseline, nor a significant interaction between compositionality and DP type in the canonical baseline. As for the interaction between compositionality and structure, a larger effect of compositionality than in the canonical baseline was found for prefield ($t = 2.92$, $p = 0.007$), scrambling ($t = 4.24$, $p < 0.001$), passive ($t = 2.61$, $p = 0.017$), nominalization ($t = 3.03$, $p = 0.007$), and *which*-question ($t = 2.63$, $p = 0.017$), while the effect was not significantly different from the canonical baseline for anaphor. The interaction between DP type and structure was not significant for any of the tested structures, but this was qualified by significant three-way interactions between structure, compositionality, and DP type for anaphor ($t = 2.19$, $p = 0.04$) and scrambling ($t = 2.63$, $p = 0.02$). In both cases, there was a stronger positive effect of compositionality with definite DPs than with indefinite DPs. The estimate of the effect of compositionality is illustrated in **Figure 10**.

In the third analysis, we fit a model to the data subset including only non-idioms in order to test whether incorporated nouns behaved differently from indefinite DPs. Structure and DP type (this time also including the third level, i.e., incorporated) were included as fixed effects. No simple effect of any of the DP contrasts was found in the canonical baseline. As for the interaction between structure and DP type, a larger contrast between definite and

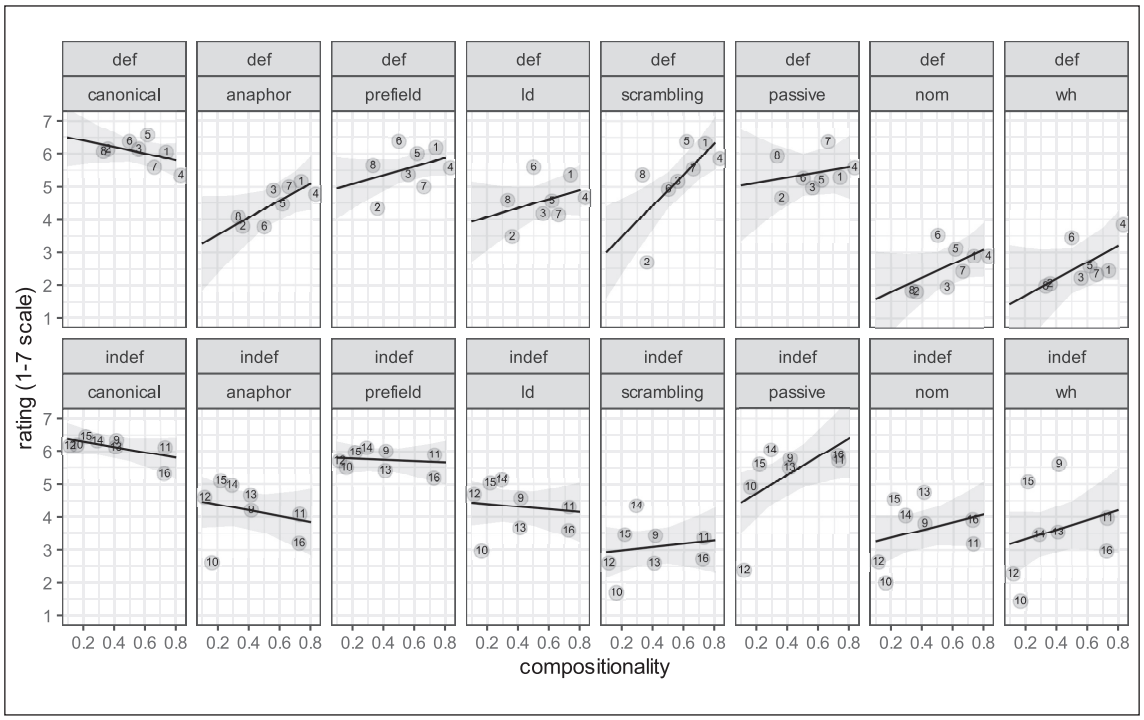

Figure 10: Estimate of the effect of compositionality (as a linear predictor) on the acceptability of the tested structures in Experiment 3 (idiom data). The circles additionally show the observed by-item means for each structure.

indefinite (towards higher ratings for definite DPs) than in the canonical baseline was found for scrambling ($t = -6.38$, $p < 0.001$). A significant interaction in the opposite direction (towards lower ratings for definite DPs) was found for *which*-question ($t = 5.01$, $p < 0.001$). As for the contrast between indefinite and incorporated, a larger contrast than in the canonical baseline (towards lower ratings for incorporated) was found for LD ($t = -5.43$, $p < 0.001$), passive ($t = -2.31$, $p = 0.03$), nominalization ($t = -2.37$, $p = 0.03$), and *which*-question ($t = -3.94$, $p = 0.001$).

All significant effects reported above according to the LMM were also significant in the CLM analysis. We summarize and discuss any additional significant effects found in the CLM but not in the LMM in Appendix B.2. It is particularly relevant for the discussion below that a significant three-way interaction between structure, compositionality, and DP type (towards a stronger positive effect of compositionality with definite DPs than with indefinite DPs) was found for anaphor and scrambling both in the LMM and in the CLM analysis. We will focus on scrambling and anaphor in the discussion in Section 5.2. In addition, this interaction was also significant for LD and nominalization in the CLM. We will remark on this in the discussion of potential directions for future research in Sections 5.2 and 5.4.

4.2.4 Discussion

Experiment 3 extended the data set to a new set of idioms and non-idioms. The main finding of Experiment 1 was confirmed: again, we observe a systematic positive linear effect of our compositionality measure on most of the tested syntactic structures. This supports the view that this is a replicable generalization.

An additional finding of Experiment 3 was that compositionality interacts with definiteness. This interaction was particularly strong for scrambling and anaphor; in these conditions, compositionality had a much stronger positive effect for definite than indefinite DPs.

When comparing the results for idioms and non-idioms, we observed that definiteness had a similar effect on both sets of expressions: for those syntactic structures in which definite non-idiomatic DPs were more acceptable than indefinite ones (in particular scrambling), we see the same pattern for idioms. For those structures in which indefinite non-idioms were more acceptable than definite ones (*which*-question and nominalization), we see the same for idioms. This adds support to the view that idioms and non-idioms are subject to similar grammatical constraints. We will discuss the definite > indefinite effect in the scrambling condition in more detail in Section 5. As for the opposite effect in the *which*-question and nominalization (indefinite > definite), we can think of the following potential reasons for it. The *which*-questions were constructed a bit differently in this experiment in order for us to be able to use the same construction for all items (with definite/indefinite/incorporated nouns): instead of a question starting with *Welche(r/s)*… 'Which…'', we used *Was für ein*… 'What kind of…'. This was potentially perceived as pragmatically odd in the conditions with definite expressions, where a referent was already introduced in the context (e.g., '*Sasha complains about his old sofa every day. Do you think he wants to get rid of it? – I don't know, what kind of couch would he throw away?'*). Similarly, for nominalization, we chose a specific way to construct it to make it work with all the DP types (using *von* 'of' rather than genitive inflection), and it is possible that this was perceived as odd with definite DPs. Even though the reasons for the definite/indefinite asymmetry in *which*-question and nominalization could have to do with the specific way we constructed these conditions, we still think it is informative that idioms and non-idioms behaved in a similar way.

While definiteness showed a clear effect on idiom and non-idioms, it cannot account for all the contrasts in the data. We still observed a gap in syntactic flexibility between definite idioms and non-idioms, and between indefinite idioms and non-idioms. In Section 5, we will discuss referentiality as a property that might provide a unified explanation for our observations, including the residual idiom/non-idiom gap.

An observation from Experiment 3 that is compatible with the idea that referentiality plays an important role concerns the items with incorporated nouns. Items with incorporated nouns are non-idiomatic (in the sense that they do not involve figurative meaning), but lack referentiality.

On average, they showed significantly lower syntactic flexibility than the other types of non-idiomatic expressions that we tested.

On a methodological note, asking two separate questions in the compositionality judgment task revealed that participants more frequently failed to find a figurative meaning for the verb than for the DP. As discussed in Sections 3.1–3.2, this can either mean that the idiom was non-compositional, or that the availability of a literal interpretation of the verb interfered[12] – this is an uncertainty that our method leaves open, and that could be addressed in future work by asking "Does the verb/DP provide its own meaning?" rather than "…its own figurative meaning"; this would also open up the possibility to have both idioms and non-idioms judged using the same task.

5. General discussion

5.1 Summary of main findings of Experiments 1–3

For German, Experiments 1 and 3 confirm a systematic effect of compositionality on syntactic flexibility. The structures that consistently showed a compositionality effect in both experiments were: prefield, scrambling, and *which*-question. Experiment 3 revealed that for scrambling, the effect is limited to idioms with definite DPs. A similar definite-indefinite asymmetry was found for idioms in the anaphor condition. Non-idiomatic VPs with an incorporated noun overall showed less syntactic flexibility than those with a definite or indefinite DP. For English, we did not find an effect of compositionality on syntactic flexibility.

5.2 Discussion of Experiments 1–3

Conceptually, compositionality has the advantage that it could in principle account for the variability found in the syntactic flexibility of idioms, and for the gap in syntactic flexibility between idioms and non-idioms. However, our results show that compositionality interacts with other factors, and that it does not affect all idioms and structures alike. We will focus on discussing the scrambling and anaphor conditions here, which both showed a consistent and interesting interaction with definiteness in Experiment 3.

How should compositionality of idioms affect the availability of scrambling? If our assumptions from 4.1.2 – that scrambling presupposes specificity of the object, and specificity presupposes compositionality – are correct, we would expect the following pattern: among non-idioms, scrambling should be more acceptable with definite than with indefinite DPs (for which a specific reading was not salient in our items) or incorporated nouns. Among idioms, scrambling should be more available for definite than indefinite DPs. Additionally, there should be an effect of compositionality within the idioms with a definite DP: the more compositional an idiom is, the easier it should be to interpret

[12] As pointed out by a reviewer, in Experiment 3 this caveat might apply to items containing semantically light/empty verbs, e.g. *die Fliege machen* (lit. 'to do the fly', fig. 'to run away').

the DP as specific, and the more acceptable scrambling should be. The pattern is summarized in **Table 1**, and it is compatible with the data obtained in Experiments 1 and 3.

	Non-idioms			**Idioms**	
	definite	**indefinite**	**incorporated**	**definite**	**indefinite**
(salient) specific interpretation	yes	no	no	if compositional	no

Table 1: Which of our item types contain an expression with a (salient) specific interpretation?

Our findings are thus compatible with the view that the variability in syntactic flexibility of both idioms and non-idioms is derivable from specificity – at least with respect to the availability of scrambling.

Let us now turn to the anaphor condition. If our reasoning in 4.1.1 is correct, we expect to find the pattern illustrated in **Table 2**: definite and indefinite non-idioms are able to introduce discourse referents, and the same should be the case for idiomatic expressions, as long as they are compositional.

	Non-idioms			**Idioms**	
	definite	**indefinite**	**incorporated**	**definite**	**indefinite**
anaphoric potential	yes	yes	no	if compositional	if compositional

Table 2: Which of our item types contain an expression with anaphoric potential?

However, this expectation is only partly corroborated: we did find a significant compositionality effect for definite DPs, but not for indefinite DPs. Furthermore, even though we found lower acceptability ratings for VPs with incorporated nouns in comparison to indefinite DPs in most of the tested structures, anaphor was not one of them, although it is the one for which we would expect an effect.

Thus, while there is a plausible explanation of our scrambling data in terms of referentiality, our anaphor results only partly align with the expectations.

Anaphor and scrambling were the structures that consistently showed a three-way interaction between structure, compositionality, and DP type – towards a more pronounced positive effect of compositionality with definite than with indefinite DPs – in both statistical analyses (LMM and CLM). How about the other structures? In the CLM analysis (see Appendix B.2), the three-way interaction reached a significant level not only for anaphor and scrambling, but also for LD and nominalization. We take this as an indication that in future research, it would be worth pursuing the question whether a similar explanation as in the case of scrambling might be applicable to further structures.

5.3 Theoretical implications

Our results contribute another piece of evidence in favor of the view that there are grammatical factors that both idioms and non-idioms are sensitive to when it comes to syntactic flexibility. In particular, we observe similar sensitivity to the definiteness of DPs contained in idiomatic/non-idiomatic VPs.

Our results furthermore show that compositionality is a factor that consistently influences the syntactic flexibility of idioms. It is therefore desirable to be able to represent compositional idioms differently from non-compositional idioms in the theoretical model, but without reducing them to completely frozen chunks (as most of the idioms we tested show at least some degree of flexibility). This is possible in Bargmann & Sailer's (2018) model by assuming that the individual words of an idiom each have their own lexical entry, but there is semantic redundancy in non-compositional idioms. In Gehrke & McNally's proposal (2019), the special behavior of non-compositional idioms follows from their limited ability to introduce discourse referents at the level of individuals.

It is interesting to note that Gehrke & McNally's (2019) proposal for determiner variability in idioms partially builds on previous analyses of incorporated noun forms. They point out that idiomatic VPs could be seen as "one extreme of a continuum" ranging from (syntactically and semantically) highly opaque V-N combinations to highly transparent ones and including both idiomatic and non-idiomatic expressions. In this sense, idioms could be considered "no different from any other combinations of words, simply more spectacular" (Gehrke & McNally 2019:807).

5.4 Outlook

In order to gain a better understanding of the results with respect to indefinite and incorporated nouns in the anaphor condition, we think that it would be useful to control the properties of the indefinite DPs more systematically (e.g., with respect to number), and to extend the set of incorporated nouns. In particular, it would be interesting to compare syntactically complex phrases lacking referentiality – for example, quantified DPs ("no book", "some books") to bare nouns in order to potentially disentangle the effect of the degree of syntactic incorporation from high/low referentiality.

Also, the distinction between incorporated/non-incorporated nouns could be made sharper in two respects. First, as mentioned in 4.1.4, we consider it as an indicator of incorporation if a singular count noun exceptionally appears without an article (e.g., *Zeitung lesen*, lit. 'read newspaper'); on the other hand, mass nouns are typically articleless in German and we therefore categorized them as regular indefinite expressions. However, the distinction between count/mass noun is not always easy to make (e.g. for *Farbe* 'color' in *Farbe bekennen*, lit. 'to profess color', fig. 'to reveal one's intentions'). Second, Frey (2015:259) suggested that even nouns with an article might be incorporated if they represent a semantically/pragmatically "typical" object of the verb – the division between typical/untypical objects could thus also be sharpened in future research. A clearer division between

incorporated/non-incorporated might shed light on the discrepancy between the expectations and findings for the anaphor condition with incorporated/indefinite nouns in Experiment 3.

We also agree with a reviewer's suggestions that it would be desirable to generally achieve a closer similarity between the idiomatic and non-idiomatic VPs by constructing pairs that are more directly matched with respect to verb and DP properties (e.g. 'kick the bucket' / 'kick the flowerpot').

Furthermore, as some of the German structures that we did not discuss in detail here show a pattern in a similar direction as scrambling, it would be worth investigating whether there perhaps is a similar relationship with compositionality and (in)definiteness as in the case of scrambling. In this context, a follow-up investigation of the informally observed linear correlations between the LD, prefield, scrambling, and anaphor conditions in Experiment 1 would also be informative.

6. Conclusion

In three experiments, we investigated potential sources of variability in the syntactic flexibility of idioms. We first reconsidered compositionality – which has been argued to influence syntactic flexibility – from a methodological perspective, aiming to avoid a potential confound of previously used methods of eliciting compositionality ratings. We indeed found a positive effect of compositionality for German, most consistently for the conditions containing movement to the prefield, scrambling, and *which*-question. We failed to find an effect for the tested English structures.

We also reconsidered compositionality from a conceptual perspective. We argued that it is a factor that can contribute to explaining why some idioms are more syntactically flexible than others, but it cannot explain similar variability among non-idioms. We have raised referentiality as a factor that could in principle provide a unified explanation. In (4), we suggested a form that an explanatory set of rules could have. (14) shows an explanation in terms of referentiality in this format.

(14) (i) Idioms that are highly referential[13] are syntactically more flexible because some syntactic modifications require referentiality.
(ii) Non-idioms that are highly referential are syntactically more flexible because some syntactic modifications require referentiality.
(iii) Idioms are less syntactically flexible than non-idioms on average because DPs contained in idioms can only be referential if the idiom is compositional.

The reasoning in (14) is in line with approaches like Bargmann & Sailer (2018) and Gehrke & McNally (2018), whose models focus on parallels between idioms and non-idioms. According to (14), it is too simplistic to say that compositional idioms are generally more flexible syntactically. Rather, the relation is indirect: syntactic flexibility presupposes referentiality, and referentiality presupposes compositionality. Thus, the more compositional an idiom is, the more sensitive it

[13] More precisely, by "idioms that are (highly) referential", we mean idioms that contain (highly) referential expressions.

can be to factors that determine the flexibility of non-idioms, like definiteness and referentiality. More compositional idioms should thus behave more similarly to non-idioms.

As a step towards evaluating this line of thought empirically, we did a follow-up experiment on German. We extended the empirical domain to a larger set of idioms and non-idioms, testing various types of DPs. Several observations support the view in (14): for non-idioms, scrambling is more acceptable with definite DPs than indefinite DPs; we found that the more compositional an idiom is, the more it approaches this pattern. Furthermore, incorporated nouns, which we assume to be a type of non-idiom lacking referentiality, were found to be relatively inflexible in our data, corroborating (14ii). On the other hand, our results for anaphors only partially align with the referentiality hypothesis and require further research. Both definite and indefinite non-idiomatic DPs have anaphoric potential; based on (14), we would expect compositional idioms to approach this pattern, but we only found this for idioms containing a definite DP.

We think that referentiality as a potential unifying source of syntactic flexibility is worth being explored further, and that our experiments provide some points of departure for this: future research could extend the empirical domain by including further types of more/less referential non-idioms, and additional judgment tasks, potentially building and improving on our compositionality task.

Data accessibility statement

The materials, collected data, and scripts used for statistical analysis are available in our Open Science Framework (OSF) repository under https://doi.org/10.17605/OSF.IO/JPUFT.

Additional files

The additional files for this article, including Appendices A, B and C, can be found at this link: DOI: https://doi.org/10.16995/glossa.8502.s1

Ethics and consent

All participants in our studies gave their informed consent before participating. According to the local regulations by the Deutsche Forschungsgemeinschaft (DFG) under which the research was conducted, questionnaire studies that are carried out with informed participants from a healthy adult population and that do not pose a risk to the participants are exempted from approval by an ethics committee.

Acknowledgements

We dedicate this paper to the memory of Gisbert Fanselow. Gisbert passed away in September 2022, while this manuscript, which he co-wrote, was under review. He is greatly missed by us and the linguistic community.

We are grateful to our editor Lyn Tieu for excellent support throughout the reviewing process and to the anonymous reviewers for their comments and questions, which helped to improve this paper. We also thank the linguistic department at the University of Potsdam for support and helpful discussion as well as Balázs Surányi and Boban Arsenijević for invaluable input at earlier stages of this research (reported in Wierzba et al. 2023), which the studies presented here build upon. Furthermore, we thank our student assistants Anna-Janina Goecke, Ulrike May, and Johannes Rothert for their help in setting up and running the online experiments. The research reported here was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287, Project C01 (during the first phase of the SFB), and by the Deutsche Forschungsgemeinschaft – SFB 632, Project A1.

Competing interests

The authors have no competing interests to declare.

References

Abel, Beate. 2003. English idioms in the first language and second language lexicon: a dual representation approach. *Second Language Research* 19(4). 329–358. DOI: https://doi.org/10.1191/0267658303sr226oa

Bargmann, Sascha & Sailer, Manfred. 2018. The syntactic flexibility of semantically non-decomposable idioms. In Sailer, Manfred & Markantonatou, Stella (eds.), *Multiword expressions: Insights from a multi-lingual perspective*. 1–29. Berlin: Language Science Press.

Bates, Douglas & Kliegl, Reinhold & Vasishth, Shravan & Harald Baayen. 2015a. Parsimonious mixed models. *arXiv.org e-print archive ArXiv*:1506.04967 [stat.ME].

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015b. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 1(67). DOI: https://doi.org/10.18637/jss.v067.i01

Bruening, Benjamin & Dinh, Xuyen & Kim, Lan. 2018. Selection, idioms, and the structure of nominal phrases with and without classifiers. *Glossa* 3(1). DOI: https://doi.org/10.5334/gjgl.288

Diesing, Molly. 1990. *The syntactic roots of semantic partition*. PhD thesis, University of Massachussets.

Everaert, Martin. 2017. Idioms: What you see is what you get? Paper presented at the syntax of idioms Workshop, Utrecht University, 20 January.

Fanselow, Gisbert. 2018. Zur Flexibilität von Idiomen im Deutschen. *Colloquia Germanica Stetinensia* 27. 115–134. DOI: https://doi.org/10.18276/cgs.2018.27-07

Fellbaum, Christiane. 2019. How flexible are idioms? A corpus-based study. *Linguistics* 57(4). 735–767. DOI: https://doi.org/10.1515/ling-2019-0015

Fodor, Janet Dean & Sag, Ivan A. 1982. Referential and quantificational indefinites. *Linguistics and Philosophy* 5(3). 355–398. DOI: https://doi.org/10.1007/BF00351459

Fraser, Bruce. 1970. Idioms within a Transformational Grammar. *Foundations of Language* 6(1). 22–42.

Frege, Gottlob. 1891. Function und Begriff. Vortrag gehalten in der Sitzung vom 9. Januar 1891 der Jenaischen Gesellschaft für Medicin und Naturwissenschaft. Jena: Verlag Hermann Pohle. DOI: https://doi.org/10.1111/j.1438-8677.1891.tb05760.x

Frey, Werner. 2004. The grammar-pragmatics interface and the German prefield. *Sprache und Pragmatik* 52. 1–39.

Frey, Werner. 2015. NP-Incorporation in German. In Borik, Olga & Gehrke, Berit (eds.), *The Syntax and Semantics of Pseudo-Incorporation. Syntax and Semantics* 40. 225–261. Leiden: Brill. DOI: https://doi.org/10.1163/9789004291089_008

Gehrke, Berit & McNally, Louise. 2019. Idioms and the syntax/semantics interface of descriptive content vs. reference. *Linguistics* 57(4). 769–814. DOI: https://doi.org/10.1515/ling-2019-0016

Gibbs, Raymond W. & Nayak, Nandini P. 1989. Psycholinguistic studies on the syntactic behaviour of idioms. *Cognitive Psychology* 21. 100–138. DOI: https://doi.org/10.1016/0010-0285(89)90004-2

Goldberg, Adele E. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224. DOI: https://doi.org/10.1016/S1364-6613(03)00080-9

Grohmann, Kleanthes K. 2000. Copy Left Dislocation. In: Billerey, Roger & Lillehaugen, Brook Danielle (eds.), *WCCFL 19 Proceedings*. 139–152. Somerville, MA: Cascadilla Press.

Heim, Irene. 1991. Artikel und Definitheit. In: von Stechow, Armin and Wunderlich, Dieter (eds.), *Handbuch der Semantik*. 487–535. Berlin: de Gruyter. DOI: https://doi.org/10.1515/9783110126969.7.487

Horvath, Julia & Siloni, Tal. 2009. Hebrew idioms: The organization of the lexical component. *Brill's Annual of Afroasiatic Languages and Linguistics* 1. 283–310. DOI: https://doi.org/10.1163/187666309X12491131130666

Horvath, Julia & Siloni, Tal. 2019. Idioms: The type-sensitive storage model. *Linguistics* 57(4). 853–891. DOI: https://doi.org/10.1515/ling-2019-0017

Karttunen, Lauri Juhani. 1969. *Problems of reference in syntax*. Indiana University.

Kratzer, Angelika. 1998. Scope or pseudoscope? Are there wide-scope indefinites? In: Rothstein, Susan (ed.), *Events and Grammar*. *Studies in Linguistics and Philosophy* 70. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-011-3969-4_8

Kuznetsova, Alexandra & Brockhoff, Per B. & Christensen, Rune H. B. 2017. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software* 13(82). DOI: https://doi.org/10.18637/jss.v082.i13

Lebani, Gianluca E., Senaldi, Marco S. G. & Lenci, Alessandro. 2015. Modeling idiom variability with entropy and distributional semantics. *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics (QITL 6)*, Universität Tübingen.

Leiner, D. J. 2019. SoSci Survey (Version 3.1.06) [Computer software]. Available at https://www.soscisurvey.de

Lenerz, Jürgen. 2001. Word order variation: Competition or cooperation? In Müller, Gereon & Sternefeld, Wolfgang (eds.), *Competition in syntax*. 249–281. Berlin: Mouton de Gruyter.

Liddell, Torrin M. & Kruschke, John K. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* 79. 328–348. DOI: https://doi.org/10.1016/j.jesp.2018.08.009

Maher, Zachary. 2013. *Opening a can of worms: Idiom flexibility, decomposability, and the mental lexicon*. BA/MA thesis, Yale University.

Modarresi, Fereshteh & Krifka, Manfred. 2021. Anaphoric reference to incorporated objects and weak definites: Does it exist? How does it work? *Talk at the Colloque international LED, Reference: (Co-)construction and use*, Université de Grenoble.

Montague, Richard. 1974. Formal Philosophy. New Haven: Yale University Press.

Müller, Gereon. 2004. Verb-second as vP-first. *The Journal of Comparative Germanic Linguistics* 7. 179–234. DOI: https://doi.org/10.1023/B:JCOM.0000016453.71478.3a

Nunberg, Geoffrey & Sag, Ivan A. & Wasow. Thomas. 1994. Idioms. *Language* 70(3), 491–538. DOI: https://doi.org/10.2307/416483

R Core Team. 2016. R: A language and environment for statistical computing. https://www.R-project.org

Schwarz, Florian. 2009. Two types of definites in natural language. PhD thesis, University of Massachussets.

Singh, Raj & Fedorenko, Evelina & Mahowald, Kyle & Gibson, Edward. 2016. Accommodating presuppositions is inappropriate in implausible contexts. *Cognitive Science* 40(3). 607–634. DOI: https://doi.org/10.1111/cogs.12260

Starschenko, Alexej & Wierzba, Marta. 2023. L-Rex: Linguistic rating experiments [software], https://github.com/2e2a/l-rex/.

Tabossi, Patrizia & Fanari, Rachele & Wolf, Kinou. 2008. Processing idiomatic expressions: effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(2). 313–327. DOI: https://doi.org/10.1037/0278-7393.34.2.313

Tabossi, Patrizia & Wolf, Kinou & Koterle, Sara. 2009. Idiom syntax: idiosyncratic or principled? *Journal of Memory and Language* 61. 77–96. DOI: https://doi.org/10.1016/j.jml.2009.03.003

Veríssimo, João. 2021. Analysis of rating scales: A pervasive problem in bilingualism research and a solution with Bayesian ordinal models. *Bilingualism: Language and Cognition* 24. 842–848. DOI: https://doi.org/10.1017/S1366728921000316

von Heusinger, Klaus. 2002. Specificity and definiteness in sentence and discourse structure. *Journal of Semantics* 19. 245–274. DOI: https://doi.org/10.1093/jos/19.3.245

Wierzba, Marta & Brown, J. M. M. & Fanselow, Gisbert. 2023. The syntactic flexibility of German and English idioms: Evidence from acceptability rating experiments. *Journal of Linguistics.* DOI: https://doi.org/10.1017/S0022226723000105

Wulff, Stefanie. 2009. Converging evidence from corpus and experimental data to capture idiomaticity. *Corpus Linguistics and Linguistic Theory* 5(1). 131–159. DOI: https://doi.org/10.1515/CLLT.2009.006