J. Linguistics **60** (2024), 431–468.
doi:10.1017/S0022226723000105

The syntactic flexibility of German and English idioms: Evidence from acceptability rating experiments[1]

MARTA WIERZBA
Universität Potsdam
J.M.M. BROWN
Universität Potsdam & Université de Lausanne
GISBERT FANSELOW†
Universität Potsdam

(Received 15 December 2020; revised 11 December 2022; accepted 4 January 2023)

It is controversial which idioms can occur with which syntactic structures. For example, can *Mary kicked the bucket* (figurative meaning: 'Mary died') be passivized to *The bucket was kicked by Mary*? We present a series of experiments in which we test which structures are compatible with which idioms in German (for which there are few experimental data so far) and English, using acceptability judgments. For some of the tested structures – including German left dislocation, scrambling, and prefield fronting – it is particularly contested to what extent they are restricted by semantic factors and, as a consequence, to what extent they are compatible with idioms. In our data, these structures consistently showed similar limitations: they were fully compatible with one subset of our test idioms (those categorized as semantically compositional) and degraded with another (those categorized as non-

† This paper is dedicated to the memory of Gisbert Fanselow, who co-wrote this paper and without whom the research reported here would not exist. Gisbert passed away in September 2023, while our manuscript was under peer-review. As Gisbert cannot now give his authorisation to publish, Gisbert's next of kin, Carola Fanselow, has given his authorisation on his behalf, and we are very grateful to her for doing this.

[1] We would like to thank the editors and the anonymous reviewers for their very helpful comments and questions. We are also particularly grateful to Balázs Surányi – at the beginning of this research project, we explored the behavior of idioms from a broader cross-linguistic perspective (including Hungarian) together, and Balázs Surányi has provided invaluable insight and contributed to shaping the experiments reported here. In the same context, we would also like to thank Boban Arsenijević for insightful discussion and for exploring idioms in Serbian. Furthermore, we thank our student assistants Anna-Janina Goecke, Ulrike May, and Johannes Rothert for their support in constructing the materials and conducting the experiments as well as Faith Chiu for her help with testing logistics. 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.

compositional). Our findings only partly align with previously proposed hierarchies of structures with respect to their compatibility with idioms.

KEYWORDS: acceptability ratings, compositionality, experiments, idioms, syntax

1. INTRODUCTION

Idioms are compatible with certain syntactic structures to varying degrees.[2] For example, according to Fraser (1970), the idiomatic verb phrase (VP) *spill the beans* (figurative meaning: 'reveal a secret') is highly flexible – e.g. it can felicitously be passivized – whereas *kick the bucket* (figurative meaning: 'die') cannot, as illustrated in (1) and (2).

(1) a. Mary spilled the beans. *canonical word order*
b. The beans were spilled by Mary. *passive*
(2) a. Mary kicked the bucket. *canonical word order*
b. *The bucket was kicked by Mary. *passive*

Fraser (1970; building on approaches by Weinreich 1969; Katz 1966) modeled the contrast in (1) and (2) by assuming that idioms have an idiosyncratic property encoding their syntactic flexibility: *spill the beans* has a [+passive] property and *kick the bucket* has [−passive]. Fraser proposed a flexibility hierarchy of idioms, ranging from flexible idioms like *spill the beans*, which can felicitously appear in a large range of structures, to highly inflexible ('frozen') ones like *kick the bucket.* Fraser also assumed a restrictiveness hierarchy of structures: for example, (a certain type of) nominalization is more restricted than passivization, i.e. compatible with a smaller set of idioms.

Later approaches (see Section 3.1) aimed at replacing Fraser's (1970) idiosyncratic features by more general linguistic properties and thus taking steps toward explaining (i) why certain structures are more RESTRICTED than others, i.e. compatible with less idioms, and (ii) why certain idioms are more FLEXIBLE than others, i.e. compatible with more structures. These are the research questions that we aim to contribute to.

With respect to question (i), we focus on the idea that certain structures are more restricted for semantic reasons. Applied to (1) and (2), the idea would be that (1b) is more acceptable than (2b), because the passive structure has a semantic restriction that is met in (1b) but not in (2b). It has been controversially discussed for many syntactic structures whether they involve such a semantic restriction or not, and our motivation in this paper is to contribute empirical data to this discussion.

For example, it has been proposed that passivization in English requires the subject of the passive sentence to be interpretable as a topic (see Nunberg et al. 1994) – e.g. a sentence like *The apple was eaten by Mary* enforces the interpretation

[2] We intend to use 'syntactic structure' here as a theory-neutral term (as far as possible), in the sense of a specification of constituent order, form, and/or hierarchy.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

that the apple is what the sentence or discourse is about. Idiomatic VPs consisting of a verb and an object provide a good test case for hypotheses of this kind, because an idiom's individual constituents do not necessarily have individual meanings. In *The bucket was kicked by Mary*, it is less clear what it would mean for the sentence or discourse to be 'about the bucket'. In our experiments, we collect acceptability ratings for sentence sets like (3) and (4):

(3)	a. Mary ate the apple.	*canonical*	*non-idiom*
	b. The apple was eaten by Mary.	*passive*	*non-idiom*
(4)	a. Mary kicked the bucket.	*canonical*	*idiom*
	b. The bucket was kicked by Mary.	*passive*	*idiom*

If it can be confirmed that passivizing an idiom as in (4) is degraded in comparison to passivizing a non-idiom as in (3), this is compatible with the hypothesis that the passive is indeed restricted by a semantic requirement (e.g. that the object needs to be interpretable as a topic). Terminology-wise, we say that the assumption that the passive requires the object to be a topic is an example of a hypothesis about as structure being 'semantically restricted' / showing 'semantic restrictiveness'. Other examples are 'German object scrambling requires the object to have a specific interpretation' and 'English clefts require the clefted constituent to have an exhaustive interpretation'. What these hypotheses have in common is the claim that not every constituent can appear in a syntactic structure, for semantic reasons. The more 'semantically restricted' a structure is, the smaller the set of constituents that are compatible with it. These examples show that the set of possible 'semantic restrictions' is heterogeneous: the restriction can be related to information-structural properties like topicality but also to other semantic properties such as specificity or exhaustivity. In Section 2, we discuss the structures tested in our experiments in more detail: we review in which way they have been proposed to be semantically restricted and discuss controversies around them. Note that comparing sentences like (3) and (4) does not tell us exactly what type of semantic restriction (topicality / specificity /…) a structure has – our goal is limited to determining whether semantic restrictiveness is empirically detectable using idioms as a test case and, if so, whether there are structures that are more restricted than others.

As for research question (ii), why certain idioms are more flexible than others, we focus on the COMPOSITIONALITY of idioms as a potential factor influencing their flexibility. We follow Nunberg et al. (1994; building on Nunberg 1977; Wasow et al. 1984) in distinguishing between two categories of idioms. First is compositional idioms that have an isomorphic mapping between parts of literal and figurative meanings: e.g. in *spill the beans*, *spill* has 'reveal' as its non-figurative counterpart, and *the beans* corresponds to 'the secret'. Thus, the meaning of the whole VP can be derived via the usual compositional mechanisms. Second, there are non-compositional idioms that cannot be broken down in this way. Their meanings cannot be derived compositionally; rather, it is related holistically to the

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

whole chunk – e.g. in *kick the bucket*, the VP as a whole means 'to die', but the parts *kick* and *bucket* do not have individual figurative meanings.[3] In our experiments, we investigate whether compositional idioms are more acceptable in various syntactic structures (passive / scrambling /…) than non-compositional ones. We therefore compare not only idioms to non-idioms but also compositional idioms to non-compositional ones. We think that this can provide an even more fine-grained estimation of a structure's semantic restrictiveness. Thus, we test the following range of conditions:

(5)	a. Mary ate the apple.	*canonical*	*non-idiom*
	b. The apple was eaten by Mary.	*passive*	*non-idiom*
(6)	a. Mary spilled the beans.	*canonical*	*compositional idiom*
	b. The beans were spilled my Mary.	*passive*	*compositional idiom*
(7)	a. Mary kicked the bucket.	*canonical*	*non-comp. idiom*
	b. The bucket was kicked by Mary.	*passive*	*non-comp. idiom*

We focus on compositionality (rather than other idiom properties like literality or predictability) because we assume that compositionality interacts with semantic restrictiveness: in a compositional idiom like (6), the phrase *the beans* can be considered to have an individual (figurative) meaning. Therefore, for a sentence like (6b), assigning a topic interpretation to *the beans* (i.e. interpreting the sentence as being about the secret that the beans figuratively refer to) is more easily conceivable than for *the bucket* in (7b). Besides passive, we test a range of further syntactic structures for which it is contested to what degree they are semantically restricted, focusing in particular on German structures for which there is not yet much empirical data.

Section 2 provides background about the syntactic structures that play a role in our experiments. Section 3 summarizes previous research. In Section 4, we present our four experiments: we provide first results for a range of German structures and we retest the flexibility of English idioms (that has been investigated in previous research).

2. Background: Controversies about structures and their restrictions

2.1 *Motivation for our selection of structures and proposed hierarchies*

Three of our four experiments involve German structures. It is interesting to focus on German for the following reasons. First, in contrast to English, there are few experimental data available on the flexibility of German idioms (see Section 3.2). Second, there are German structures for which it is particularly contested to what extent there are limitations in the compatibility with idioms and whether these can

[3] For further discussion of compositionality in the context of idioms and phrasal units, see also e.g. Goldberg (2006), Jackendoff (2008), and references therein.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

be attributed to semantic restrictions: German prefield fronting, scrambling, and left dislocation (LD).

A hierarchy similar to Fraser's (1970) has been proposed for German by G. Müller (2000, 2019). According to G. Müller (2000), prefield fronting is one of the least restrictive structures in German, followed by passivization, *wh*-movement (in particular, *which*-questions), and finally LD. G. Müller (2019) additionally proposes that scrambling is more restrictive than *wh*-movement. Like Fraser's proposal, G. Müller's hierarchy is also based on the compatibility of the structures with different groups of idioms.[4] The structures are illustrated in (8) for a non-idiom; examples from the literature containing idioms are provided throughout this section.

(8) a. Den Apfel hat Maria gegessen. *prefield*
the apple has Maria eaten
'Maria has eaten the apple.'

b. Der Apfel wurde (von Maria) gegessen. *passive*
the apple was by Maria eaten
'The apple was eaten (by Maria).'

c. Welchen Apfel hat Maria gegessen? *which-question*
which apple has Maria eaten
'Which apple has Maria eaten?'

d. Den Apfel, den hat Maria gegessen. *left dislocation*
the apple PRONOUN has Maria eaten
'As for the apple, Maria ate it.'

e. Maria hat den Apfel wahrscheinlich gegessen. *scrambling*
Maria has the apple probably eaten
'Maria has probably eaten the apple'

In our fourth experiment, we also test English structures. The experiment provides a link between our studies and previous experimental research on idiom flexibility, which has focused on English. Experiment 4 also serves to test the empirical adequacy of Fraser's (1970) hierarchy, according to which English nominalization without 'of' is one of the least restricted operations, passivization is more restricted, followed by nominalization with 'of', and clefting as the most restricted structure.

(9) a. I'm talking about Mary's spilling the beans. *nominalization without 'of'*
b. The beans have been spilled by Mary. *passive*
c. I'm talking about Mary's spilling of the beans. *nominalization with 'of'*
d. It is the beans that Mary spilled. *cleft*

[4] G. Müller (2019: 439) refers to the relevant idiom property as 'opacity'; similar to Fraser's (1970) model, this property directly encodes compatibility with syntactic structures in an implicational manner: 'If an idiom α dominates an idiom β on the opacity scale, and transformation δ can affect α, then δ can also affect β'.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

In Sections 2.2–2.9, we provide background on the structures that we test. We aim to provide examples from the literature showing that their compatibility with idioms is controversial and give illustrative examples of hypotheses about semantic restrictiveness.

2.2 *German prefield*

Prefield fronting refers to placing a constituent in the sentence-initial position preceding the finite verb in German main clauses. G. Müller's (2000) assumption that it is one of the least restrictive structures in German is in line with the fact that even examples involving a non-compositional idiom in the prefield have been judged as felicitous in the literature. For example, the sentence in (10a) from Ackerman & Webelhuth (1993) is discussed as a grammatical example by Nunberg et al. (1994); see also Webelhuth & Ackerman (1999).

(10) a. Den Vogel hat Hans abgeschossen.
the bird has Hans shot
'Hans stole the show.'

b. Den Garaus hat Hans dem Kollegen gemacht.
the GARAUS[5] has Hans to.the colleague done
'Hans killed the colleague.'

However, the status of sentences like (10b) is not completely uncontested. Nunberg et al. (1994: 512) report 'a good deal of variability regarding these judgments', and disagreements are also sporadically found in the literature. For example, (10b) is judged as well-formed by Fanselow (2004) but not by Frey (2004a).

Examples of theoretical approaches assuming prefield fronting to be semantically restricted are e.g. Fanselow (2004), who discusses a connection between the prefield position and focus or topic interpretation, and Frey (2004a, 2010), who proposes a contrastive or emphatic interpretation of objects in the prefield, which should not be possible if the object is part of a non-compositional idiom. On the other hand, there are also analyses according to which prefield fronting does not have a semantic effect but rather is purely formal; this is the explanation that Nunberg et al. (1994) propose for the high acceptability of idiom parts in the prefield; see also e.g. Fanselow & Lenertová (2011).

2.3 *German scrambling*

Scrambling refers to placing a constituent toward the left periphery of the sentence but without crossing the finite verb in main clauses or the complementizer in

[5] Glosses in small caps are used for unical elements, i.e. expressions that practically only appear within the idiom and that do not have a straightforward literal translation (see Soehn 2006: §2.2.3).

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

subordinate clauses. In G. Müller's (2019, 2020) hierarchies, German scrambling is assumed to be among the most restricted structures, even more restricted than *which*-questions.

A study reported in Fanselow (2010) suggested that scrambling is acceptable at least for idiom parts that are definite (see Section 3.3). This view is not shared by everyone; a negative judgment for scrambling of a definite idiom part is e.g. reported by S. Müller (2010: 610), see (11).

(11) a. … dass er dem Mann den Garaus gemacht hat.
that he to.the man the GARAUS done has
'…that the killed the man.'
b. … *dass er den Garaus dem Mann gemacht hat.

There are approaches according to which scrambling moves specific (Diesing 1990) or topical (Frey 2004b) objects toward the left periphery of the clause; this would predict that scrambling of idiom parts should be limited.

2.4 *German left dislocation*

German LD is a structure in which a constituent is placed in sentence-initial position and followed by a pronominal element. Cardinaletti (1986: 226, endnote 24) and Grohmann (2000: 144) argued that prefield fronting and LD involve similar syntactic structures and report (12) and (13) as acceptable, respectively.

(12) Den Garaus, den will er mir machen.
the Garaus GARAUS wants he to.me do
'He wants to kill me.'

(13) Den Kopf, den hat Alex der Maria gestern verdreht.
the head PRONOUN has Alex to.the Mary yesterday turned
'Alex sweeped Mary off her feet yesterday.'

In contrast, negative judgments are provided by Jacobs (2001: 677, endnote 33), who notes that 'idiom chunks (which are clearly non-referential) cannot be left-dislocated', and by G. Müller (2000), who assumes that LD belongs to the most restricted structures.

An example of a theoretical approach that assumes prefield fronting to be semantically restricted is Frey's (2004c) analysis of LD as a topic-marking structure.

2.5 *Pronominalization*

Pronominalization is another case in which the degree of semantic restrictiveness is controversial and thus worth testing. Nunberg et al. (1994) argue that the antecedent of a pronoun needs to (individually) refer to something, which is the case for parts of

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

compositional idioms like *keep tabs on* (figurative meaning: 'to monitor someone'). They judge (14) as felicitous. The example stems from Bresnan (1982), who judged it as ungrammatical and argued that idiom parts cannot be antecedents for pronouns.

(14) Although the F.B.I. kept tabs on Jane Fonda, the C.I.A. kept them on Vanessa Redgrave.

In contrast to German prefield fronting, scrambling, and LD, the potential restriction does not have to do with information-structural properties like topicality. Rather, the question is whether an idiom part like *tabs* it is able to introduce a referent that can be picked up by a pronoun like *them* in (14).

2.6 *Passivization*

In Experiments 3 and 4, we also test passivization. It has been argued that German passive is relatively unrestricted semantically – Nunberg et al. (1994) proposed that this is the explanation for the observation that in German more idioms can undergo passivization than in English, which requires the promoted argument to be a topic (as discussed in Section 1). This is in line with G. Müller's (2000) placement of German passive toward the unrestricted end of the hierarchy.

2.7 *Nominalization*

Nominalization is a structure that we would not expect to be restricted for semantic reasons. It is thus interesting and worth testing that one type of it (nominalization with 'of') has been claimed to be one of the most restricted structures in English by Fraser (1970), while another type (without 'of') has been claimed to be one of the least restricted structures.

2.8 Which-*questions*

Which-questions are a structure that we definitely expect to be semantically restricted and to not be compatible with all idioms. A sentence like *Which bucket did Mary kick?* is associated with question semantics along the lines of 'For which x, x being a bucket, is it true that Mary kicked x'?[6] For the question to make sense semantically, it is necessary that *bucket* has a meaning (in particular, that it denotes a property); otherwise, *x is a bucket* could not be evaluated. If semantic restrictiveness can be detected using test sentences with idioms at all, we have the clear expectation that *which*-questions will be restricted. In that sense, *which*-questions serve as a methodological sanity check; it can help to make sure that method is sensitive enough to detect semantic restrictiveness at all.

[6] Formally, the question denotation can be modeled as the set of all true propositions 'Mary kicked x', where x is a bucket (Karttunen 1977).

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

2.9 *Cleftlike structures*

Another structure that we would expect to be highly restricted and that can thus also serve as a methodological sanity check are English clefts. In Fraser's (1970) hierarchy, clefts are a highly restricted structure. We discuss cleft semantics in Section 4.4 in the context of our Experiment 4 on English.

2.10 *Aims and limitations*

We have mentioned a range of theoretical approaches as examples of assumptions about semantic restrictiveness of structures. We want to reiterate that the focus of this paper is not on evaluating the models' assumptions about the exact nature of the potential semantic restrictions, i.e. we do not aim to answer questions like, Is prefield fronting restricted because it involves topicality or contrast?

What the various approaches crucially have in common is that they all make empirical predictions concerning the compatibility of certain syntactic structures with certain idioms. In view of the controversies concerning the acceptability of the discussed examples, we think it is useful to start at the basic level of determining which of these claims are descriptively accurate: Can we e.g. find evidence that prefield fronting is degraded with idioms, only degraded with some idioms (potentially non-compositional ones), or compatible with all idioms? What about other structures like scrambling and LD: Can we find evidence that they are more restricted than prefield fronting in this respect? Answering these empirical questions contributes to answering our two research questions: (i) Are some structures more restricted than others due to semantics, and (ii) Are some idioms more flexible than others due to compositionality? In Section 4.6, we give an outlook on the question of how a theoretical model's exact predictions could be investigated in more depth in future work.

3. Previous research

3.1 *Previous theoretical approaches*

The line of thought – that the acceptability of sentences containing idioms depends on how semantically restricted the structure is and how compositional the idiom is – can already be found in Nunberg et al. (1994), even though this argument is not the main focus of their paper. As discussed in Section 1, they argued that English passivization is semantically restricted because it involves topicality. For further approaches that take up and discuss this idea, see also Kay & Sag (2014) for English and Bargmann & Sailer (2018) for German. Bargmann & Sailer (2018) argue against Nunberg et al.'s (1994) categorical split between compositional and non-compositional idioms, but they agree with the idea that there is a crucial interaction between semantic properties of the idioms and semantic (language-specific) restrictions of syntactic structures.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

3.2 *Previous experimental research on English*

Gibbs & Gonzales (1985) conducted experiments in which they first collected ratings of syntactic flexibility or frozenness and then collected processing times. In the frozenness pre-test, participants were asked to rate how similar the meaning of sentence pairs is, where one sentence contained an idiom and the other a non-idiomatic paraphrase:

(15) a. Her father's laying down the law prevented her from going to the dance.
b. Her father's giving strict orders prevented her from going to the dance.

The rating was intended to reflect the extent to which the idiomatic reading stays intact even when the syntax is modified. Gibbs & Gonzales tested five syntactic structures: nominalization without 'of', adverb insertion,[7] particle movement, passive, and nominalization with 'of'. Low average similarity ratings across all tested structures were taken as an indicator of a 'frozen' idiom and high average similarity ratings as an indicator of a flexible idiom. They found that nominalization with 'of' received the highest similarity ratings, whereas the other structures did not differ significantly from each other. A by-item analysis showed a gradient continuum of syntactic flexibility among idioms. The flexibility estimate for each idiom gained in this first experiment then served as a basis for follow-up experiments on the processing of idioms. In the follow-up experiment, they found, among other results, faster reaction times to frozen idioms than to flexible ones in a lexical decision task.

In another influential study, Gibbs & Nayak (1989) investigated whether the variability in syntactic flexibility found by Gibbs & Gonzales (1985) can be attributed to the semantic compositionality of the idioms. They collected judgments of compositionality in a first experiment and then tested whether this factor interacted with syntactic flexibility in a second experiment. Participants were asked to divide a set of idioms into three categories: 'normally decomposable' ones, in which each word makes a 'unique contribution to the phrase's nonliteral interpretations' (Gibbs & Nayak 1989: 108); 'abnormally decomposable' ones, where the relation between the literal and non-literal interpretation is less direct; and 'non-decomposable' ones. For the experiment on syntactic flexibility, Gibbs & Gonzales's (1985) method was adopted. They did not replicate Gibbs & Gonzales's (1985) finding of highest ratings for nominalization with 'of'; instead, nominalization without 'of' and adverb insertion received significantly higher similarity ratings than all other structures across all idioms. There was a significant interaction with compositionality: 'non-decomposable' idioms received lower ratings for adjective insertion and passive than the other idiom groups. A significant influence of compositionality was also found in a third experiment on pronominalization:

[7] As pointed out by a reviewer, the availability of adverb insertion depends on the type of adverb. We do not discuss adverbs in more detail here because they are not directly relevant to our experiments.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

'normally decomposable' idioms received higher similarity ratings than the other categories.

Gibbs & Gonzales's (1985) and Gibbs & Nayak's (1989) studies were influential with respect to the methods that they introduced for collecting compositionality categorizations and flexibility ratings. In both studies, variability between the idioms with respect to syntactic flexibility and between structures with respect to restrictiveness was found, but the findings were not consistent with respect to the question of which structure is the most restricted one.

3.3 *Previous research on German*

For German, fewer experimental data on syntactic flexibility of idioms are available than for English. Soehn (2006), Bargmann & Sailer (2018), Fanselow (2018), and Fellbaum (2019) discuss corpus and web examples that provide the valuable insight that for most syntactic modifications, it is possible to find examples involving idioms, even with ones that are intuitively non-compositional. However, there are some challenges in interpreting the corpus data: without the possibility of comparing minimal pairs, it is difficult to assess which of the occurrences can be considered as evidence for syntactic flexibility and which cases are instances of deliberate bending of linguistic rules for rhetorical purposes.

An acceptability rating experiment on German idioms is reported by Fanselow (2010), who tested whether idioms parts can be scrambled. No significant difference was found between sentences with and without scrambling when the scrambled element was definite (5.8 vs. 5.4 on a 7-point scale), whereas a significant difference was found when it was indefinite (5.9 vs. 4.4). These results tentatively point toward the conclusion that scrambling is not so much restricted in terms of semantics or compositionality but rather by the formal property of (in)definite marking.

A detailed experimental investigation comparing a range of syntactic structures has not been done yet for German.

3.4 *Methodological remarks*

Because one of our goals is to investigate the properties of various syntactic structures, it is crucial to employ a method that allows to quantify as exactly as possible whether a certain structure is less acceptable with idioms in comparison to non-idioms.

In the studies reported in Section 3.1, the results for the tested structures varied between experiments. Whereas Gibbs & Gonzales (1985) found the highest mean rating for nominalization with 'of' in their Experiment 1 (5.58 on a 7-point scale) and the lowest for adverb insertion (4.22), Gibbs & Nayak (1989) report the opposite for their Experiment 2: nominalization with 'of' received the lowest mean rating (4.44) and adverb insertion received the highest (5.28), although the same methodology was used.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

Furthermore, Gibbs & Nayak's (1989) was replicated for Italian by Contrary to Tabossi et al.'s (2008) expectations and to Gibbs & Nayak's findings for English, adverb insertion was the structure that showed the largest difference between compositional and non-compositional idioms in Tabossi et al.'s (2009) study, while no differences in this respect were found between the other tested syntactic structures (adjective insertion, passive, and LD). Tabossi et al.'s (2008) findings thus further support the view that there is a systematic relation between compositionality and syntactic flexibility, but their findings deviate from Gibbs & Nayak's findings with respect to which structures show the strongest effect of compositionality, although the same methodology was used and no difference was expected between the languages on theoretical grounds.

These studies had a different focus; thus, the deviations are not crucial for the main conclusions that the authors of these studies draw, e.g. concerning correlations between compositionality and processing measures. For our purposes, however, it is crucial to get reliable estimates of the acceptability of each structure; thus, it is important to consider what could have caused the deviations between experiments.

The deviations might have several reasons. First, the employed methodology requires to choose a paraphrase for each idiom. This might introduce a confound: when participants are asked to judge how similar e.g. a sentence like 'the law was laid down' is to the paraphrase 'strict orders were given' (Gibbs & Gonzales 1985: 245), the judgment arguably depends not only on the passivizability of 'lay down the law' but also on the passivizability of 'give strict orders'; a different paraphrase might lead to a different result. Second, it is possible that even though the ratings are intended to reflect whether the expression retains its idiomatic meaning in the given syntactic structure, the ratings might also be influenced by other factors like plausibility or complexity – for example, it might be more difficult to judge whether two sentences have the same meaning when the sentence is harder to process, which might result in a lower rating. Because no minimal pairs or sets of items were used across structures, e.g. 'They will lay the law down / give strict orders if the party gets too wild' for particle movement vs. 'Her father's laying down the law / giving strict orders prevented her from going to the dance' for nominalization (Gibbs & Gonzales 1985: 245), differences between the conditions might result from such non-syntactic factors.

We therefore employ a different method, namely acceptability ratings, as also used by Maher (2013).[8] This exempts us from the need to provide a paraphrase for the idioms and thus eliminates the risk of confounds connected to this. Furthermore, we aim to construct minimal pairs that only differ in the syntactic manipulation of interest. Using acceptability instead of similarity ratings also comes with the advantage that it allows adding an important baseline: in addition to comparing

[8] When we started conducting the first experiments reported here, we were not aware of Maher's (2013) experiments yet; the decision to use acceptability ratings was taken independently. The focus of Maher's experiments was also on the syntactic flexibility of idioms but from a different angle: Maher compared participants' reaction to syntactic modification of real vs. invented idioms, which was a replication of a study by Tabossi et al. (2009) on Italian.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

different types of idioms, we can also compare them directly to non-idioms. We also include a baseline with canonical word order. The baselines help to quantify exactly to what extent any observed differences are due to the specific interaction between idiomaticity and syntax that we are interested in.

A further relevant finding, reported by Tabossi et al. (2009), is that providing a pragmatically suitable context significantly raises the ratings for syntactically modified idioms (except when the modification violates a grammatical requirement). We also provide contexts to ensure that a sentence is not just rejected because pragmatic motivation to use this structure at all is lacking.

We employ this methodology throughout our four experiments.

4. Experiments

In Experiments 1 and 2, we investigate the syntactic flexibility of idioms in German structures, which have been proposed as differing in their semantic restrictiveness and, as a consequence, in their compatibility with idioms: prefield fronting, LD, and scrambling. For comparison, we also test pronominalization. In Experiments 3 and 4, we extend the empirical range to further structures.

The main goal of our experiments is to investigate (i) whether we can identify structures that are more semantically restricted than others, which we assume to be detectable in terms of more limited compatibility with idioms, and (ii), whether some idioms are more flexible than others due to compositionality.

To achieve this, we look at relative differences between conditions, e.g. are non-compositional idioms less acceptable than compositional ones in the prefield position / LD / scrambling, and is this acceptability difference larger than in a sentence with canonical word order? This allows us to conclude whether speakers systematically perceive contrasts between different types of idioms with respect to their syntactic flexibility (i.e. which syntactic structures they are compatible with) as well as whether speakers perceive contrasts between different types of syntactic structures with respect to their restrictiveness (i.e. which idioms they are compatible with).

All items, results, and analysis scripts are available in our Open Science Framework (OSF) repository under https://osf.io/b496a.

4.1 *Experiment 1 (German, first set of structures)*

4.1.1 *Participants and procedure*

Participants were recruited at the University of Potsdam using Sona Systems. Prescreening filters ensured that all participants were native speakers of German. Most of the participants in Experiments 1–4 were bachelor's degree students. They received payment or course credit; 41 speakers took part.

A web-based questionnaire was set up using SoSciSurvey (Leiner 2018); 121 stimuli were presented to each participant (90 critical items; the remaining

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

stimuli partly stemmed from an unrelated study and partly served to check some caveats concerning participants' general reaction to idiomatic materials; see Appendix A[9]). Each stimulus was a dialog consisting of a context question and a response or reaction, separated by a line break. Participants were instructed to judge how acceptable the response or reaction was in the given context. The instruction was to decide whether the sentence could be uttered in this form in a (possibly informal) conversation. Participants were given a scale from 1 (very unacceptable in this context) to 7 (very acceptable in this context).

4.1.2 *Design and materials*

Three factors were manipulated in this study in a 3 × 2 × 5 design. The first factor was COMPOSITIONALITY of the VP with three levels: non-idiomatic, compositional idiom, and non-compositional idiom. The second factor was CONTEXT with two levels: a broad focus context and a polarity focus context (illustrated below). The third factor was syntactic STRUCTURE with five levels: canonical word order (baseline), fronting to the prefield, LD, scrambling, and pronominalization or anaphor. We describe the factors and their levels in in more detail below.

We selected German VP idioms from a corpus-based collection by Jan-Philipp Soehn. The idioms are listed in Table 1. All consist of a definite determiner phrase (DP) and a verb. Two of us (native speakers of German) categorized them independently for compositionality based on our intuition, following the criterion of

Category	Idiom	Translation (literal)	Translation (figurative)
non-comp.	das Handtuch werfen	'to throw the towel'	'to give up'
non-comp.	den Garaus machen	'to make the GARAUS'	'to kill'
non-comp.	das Zeitliche segnen	'to bless the temporal'	'to die'
non-comp.	den Löffel abgeben	'to give away the spoon'	'to die'
non-comp.	die Leviten lesen	'to read the LEVITEN'	'to scold'
non-comp.	die Sau rauslassen	'to release the pig'	'to party excessively'
comp.	das Kriegsbeil begraben	'to bury the hatchet'	'to end a conflict'
comp.	den Braten riechen	'to smell the roast'	'to become suspicious'
comp.	den Faden verlieren	'to lose the thread'	'to get lost (e.g. in a conversation)'
comp.	den Laufpass geben	'to give the LAUFPASS'	'to break up'
comp.	den Tiefpunkt erreichen	'to reach the lowest point'	'to be in the worst possible situation'
comp.	das Eis brechen	'to break the ice	'to reduce the social tension'

Table 1
German idioms used in Experiments 1–3.

[9] Appendices A–C with additional information are available as supplementary files to this article.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

whether both the verb and the DP have their own individual figurative meaning that combine compositionally to form the figurative meaning of the VP. We selected 12 idioms with congruent annotation, aiming at choosing maximally clear cases: six idioms that we categorized as compositional (i.e. as having individual figurative meanings for the DP or verb) and six that we both categorized as non-compositional.

Note that the notion of semantic restrictiveness that we are interested in is not included explicitly as a factor in our experiment. We are able to draw conclusions about semantic restrictiveness by considering the factors STRUCTURE and COMPOSITIONALITY. Our assumption is that if a syntactic structure is more semantically restricted than another, this should be reflected in a certain kind of interaction with our idiom categories: a highly restrictive structure should be compatible with fewer idioms (by assumption, only with compositional ones). For this to work, it is crucial that the same idiom groups are used in all critical conditions (i.e. that minimal pairs or sets of items are used) and that one of the categories is overall more likely to contain compositional idioms than the other – thus, our intuitive categorization should provide a viable proxy of compositionality for our purposes, even if the distinction between the categories is not categorical and completely clear-cut. However, it is important to note that the intuitive categorization comes with certain caveats – we return to this issue in Section 4.5, where we take a closer look at individual idioms, and in the outlook on directions for further research in Section 4.7.

We added six non-idiomatic VPs: *den Bus verpassen* 'to miss the bus', *den Manager verärgern* 'to upset the manager', *die Fenster putzen* 'to clean the windows', *den Hausschlüssel verlieren* 'to lose the housekey', *den Rasen mähen* 'to mow the lawn', and *die Anlage ausmachen* 'to turn off the stereo'.

The same 12 idioms and six non-idioms were used throughout Experiments 1–3.

The target sentences were presented in one of two different contexts. CONTEXT was a between-subjects manipulation: each participant either saw all items in the first type of context or all items in the second type of context. The first context type always contained a *why*-question, inducing broad focus in the answer sentence. The second context type contained a *yes / no*-question, inducing polarity focus in the answer sentence. The VP in the target sentence was discourse-given in the latter context type by virtue of a synonym used in the *yes / no*-question. The two context types are illustrated in (16).

(16) a. broad focus context:
Maria und Peter haben doch immer gegen die ungerechte Behandlung der Auszubildenden gekämpft. Warum habe ich in letzter Zeit nichts mehr darüber gehört? *'Mary and Peter always used to fight against the unfair treatment of the trainees. Why haven't I heard about that lately?'*
– Die beiden haben wohl das Handtuch geworfen!
'The two of them have apparently thrown in the towel!'

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

b. polarity focus context:
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?'*
– Nein, die beiden würden bestimmt nie das Handtuch werfen!
'No, the two of them would definitely never throw in the towel!'

One purpose of systematically controlling the context is to get an impression of whether decreased acceptability can be alleviated by providing a suitable context.

In addition, we hypothesize that the context manipulation can potentially provide a further way to identify semantically restricted structures. The two contexts induce different information-structural effects in the target sentence. If it is the case that some of the tested syntactic structures involve a topical and/or contrastive interpretation of the object (as it has been proposed for prefield fronting, scrambling, and LD), we might see an informative difference between the context types. In the broad focus context, the VP provides new information. Thus, the context is not compatible with interpreting the VP or a part of it as given information or topic. The polarity focus context, on the other hand, leaves more scope for interpretation in this respect. Here, the denotation of the VP is discourse-given by virtue of the synonym in the question (e.g. give up – throw in the towel). Interpreting the VP as topical and/or contrastive would require a certain amount of accommodation but would not be at odds with the context: the reader could construe an implicit contrast like 'As for throwing in the towel, the two of them would never do that (in contrast to other activities)'. For non-idioms and compositional idioms, the possibility of interpreting the direct object on its own as a contrastive topic could also be available at least to some extent ('As for the ice, the two of them would never break that (in contrast to other things)', while this should not be possible with non-compositional idioms. The predictions are spelled out in more detail in Section 4.1.3.

We constructed five syntactic variants of each of the 18 items (the 12 idioms and six non-idioms).[10] This is illustrated in (17) for the polarity focus context. The structure of the anaphor condition is constructed similarly as Nunberg et al.'s (1994: 502) example *We thought tabs were being kept on us, but they weren't,* which is a modification of one of Bresnan's (1982) examples.

(17) 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!'

[10] A further property that we controlled for was whether the subject of the target sentence was a pronoun or a full DP. Half of the participants saw pronouns or DPs, respectively. This manipulation was intended to provide a first exploratory look at predictions of specific models (in particular, the distinction between pronoun and DP plays a role in Frey 2004a, 2010), which are, however, not discussed within this paper for reasons of space.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

b. prefield:
Nein, das Handtuch würden die beiden bestimmt nie werfen!
no the towel would the two definitely never throw
c. left dislocation:
Nein, das Handtuch, das würden die beiden bestimmt nie werfen!
no the towel PRON would the two definitely never throw
d. scrambling:
Nein, die beiden würden das Handtuch bestimmt nie werfen!
no the two would the towel definitely never throw
e. anaphor:
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!'

The stimuli were distributed such that every participant saw every item in all levels of the factors STRUCTURE and COMPOSITIONALITY (only CONTEXT was manipulated between subjects). The motivation for this was that we assumed that there might be variation between participants with respect to how familiar they are with each idiom and how acceptable they find it in general; by collecting ratings for each idiom in all conditions, we can be more confident that any differences that we find can be attributed to the syntactic manipulation. This decision led to a high number of critical items that each participant saw. To limit the overall length of the experiment, we included only a relatively small number of stimuli that were not part of the critical items described above. These stimuli partly served to check some potential caveats concerning judgments of idiomatic expressions, e.g. the question of whether participants tend to generally tolerate grammatical violations with this type of material. The design and results of these additional stimuli are reported in Appendix A.

4.1.3 *Hypotheses*

Our goal is to test whether the factors COMPOSITIONALITY and CONTEXT have a systematic influence on the acceptability of sentences containing idioms. In particular, we want to know whether they show an interaction with the factor STRUCTURE. The motivation to focus on interactions in the analysis as well as potential concerns and alternatives to this analysis are discussed in Appendix B1. The following hypotheses were formulated prior to collecting data.

4.1.3.1 COMPOSITIONALITY–STRUCTURE HYPOTHESIS

If syntactic flexibility depends on compositionality, we should see an interaction between the factors COMPOSITIONALITY (in particular, we would expect to see an

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

effect when comparing the following levels: compositional vs. non-compositional idiom) and STRUCTURE (canonical baseline vs. each marked structure). If the marked (non-canonical) structures are semantically restricted, they should show a different idiom behavior than the canonical baseline: they should be compatible with a smaller subset of idioms, i.e. the acceptability gap between compositional and non-compositional idioms should be larger in comparison to the baseline (toward lower acceptability for non-compositional idioms in marked structures). We additionally test whether the marked structures differ from each other in this respect in a post hoc analysis. Contrasts in the post hoc comparisons would indicate that the structures are semantically restricted to different degrees.

4.1.3.2 CONTEXT–STRUCTURE HYPOTHESIS

If the marked syntactic structures are semantically more restricted than canonical word order, and if the semantic restriction has to do with information-structure, we would also expect a certain interaction between CONTEXT and STRUCTURE. More specifically, since the polarity focus context was constructed to facilitate a marked semantic interpretation (contrastive or topical), which has been argued to potentially play a role for prefield fronting, LD, and scrambling, we could see a CONTEXT × STRUCTURE interaction when comparing these three structures to the baseline. For anaphor, on the other hand, context is not expected to play a role, if the above reasoning is correct.

4.1.4 *Results*

The results of Experiment 1 are summarized in Figure 1 and Table 2.

For statistical analysis, the factor CONTEXT was sum-coded. For the factor COMPOSITIONALITY, forward-difference coding was used (i.e. the level compositional idiom is compared to non-idiomatic, and non-compositional idiom is compared to

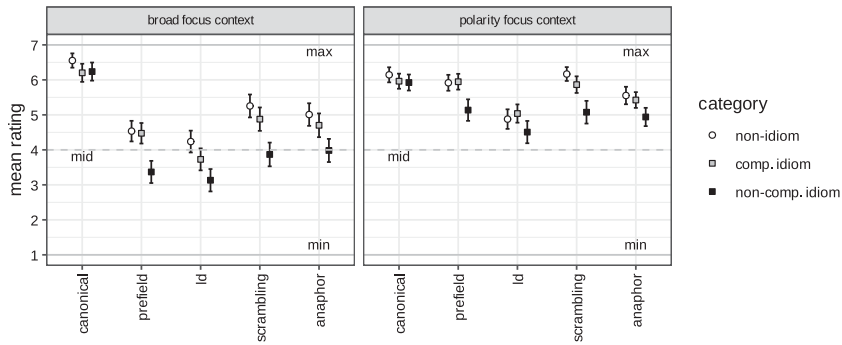

Figure 1
Experiment 1 – mean ratings (close-ended 1–7 point scale; error bars represent 95% confidence intervals).

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

Context	structure	Non-idioms	Comp. idioms	Non-comp. idioms
broad focus	canonical	6.55 (1.41)	6.20 (1.41)	6.24 (1.11)
broad focus	prefield	4.54 (1.60)	4.47 (1.60)	3.37 (1.73)
broad focus	LD	4.24 (1.70)	3.73 (1.71)	3.13 (1.74)
broad focus	scrambling	5.25 (1.78)	4.88 (1.82)	3.87 (1.85)
broad focus	anaphor	5.01 (1.76)	4.70 (1.85)	3.98 (1.80)
polarity focus	canonical	6.14 (1.26)	5.96 (1.26)	5.92 (1.34)
polarity focus	prefield	5.92 (1.33)	5.95 (1.32)	5.14 (1.80)
polarity focus	LD	4.88 (1.64)	5.04 (1.53)	4.51 (1.87)
polarity focus	scrambling	6.17 (1.16)	5.86 (1.39)	5.08 (1.90)
polarity focus	anaphor	5.55 (1.46)	5.42 (1.32)	4.94 (1.53)

Table 2
Results of Experiment 1 – mean ratings (standard deviations).

compositional idiom). For the factor STRUCTURE, treatment coding with canonical as the baseline was used (all other structures are compared to canonical). This also means that the model output for CONTEXT, COMPOSITIONALITY, and CONTEXT*COMPOSITIONALITY is to be interpreted as simple effects or interactions, i.e. effects within the baseline level of structure (canonical word order).

A linear mixed model was used for inferential statistical analysis.[11,12] For our research question, it is relevant which interactions with STRUCTURE are significant, and we therefore only report on these below; the simple effects or interactions are not informative with respect to our hypotheses. The exact specification of the models and the full output can be found in Appendix B2.

We found a significant two-way interaction between CONTEXT and STRUCTURE for each level of structure in the following direction: the acceptability difference between the canonical baseline and the other structures was reduced in the polarity context in comparison to the broad focus context for prefield ($t = -8.40$, $p < 0.001$), LD ($t = -3.78$, $p < 0.001$), scrambling ($t = -6.07$, $p < 0.001$), and anaphor ($t = -4.05$, $p < 0.001$). We also found a significant two-way interaction between COMPOSITIONALITY and STRUCTURE for each level of structure when comparing compositional idiom to non-compositional idiom: the acceptability difference between these two idiom categories was larger than in the canonical baseline for prefield ($t = -5.10$, $p < 0.001$), LD ($t = -3.63$, $p < 0.001$), scrambling ($t = -2.74$, $p = 0.01$), and anaphor

[11] The models for our hypothesis tests were fit following the recommendations for identifying parsimonious models by Bates et al. (2015a) and using the R packages lme4 and lmerTest (R Core Team 2016, Bates et al. 2015b; Kuznetsova et al. 2017).

[12] In addition to the linear mixed models reported here (using the original 1–7 ratings), we also ran linear mixed models using *z* scores and cumulative link models using the R package ordinal (Christensen 2019) to make sure that our conclusions are not based on artifacts of the analysis method. The detailed results of the latter two types of models can be found in the OSF repository. In the result sections of Experiments 1–4, any deviations between the models with respect to the significance of effects relevant for the evaluation of our hypotheses (interactions with STRUCTURE) are noted.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

($t = -3.88$, $p < 0.001$). We did not find significant interactions between COMPOSITIONALITY and STRUCTURE when comparing non-idiomatic to compositional idiom.[13] We did not find significant three-way interactions.

Besides the planned contrasts concerning the factor STRUCTURE (comparing canonical to all other levels), we also conducted a post hoc analysis to test if the structures prefield, LD, scrambling, and anaphor differ from each other with respect to the interaction with the factor COMPOSITIONALITY. The goal of these additional tests is to check to what extent our results are compatible with the proposed hierarchies holding between syntactic structures. The post hoc analysis was done by running models in which another level of the factor structure was set as the baseline. Bonferroni correction was used to compensate for the higher likelihood of erroneous inferences in multiple testing. None of these additional pairwise comparisons was significant. Numerically, the largest differences were found for prefield vs. LD and prefield vs. anaphor when comparing compositional idioms to non-compositional idioms. The detailed results of the post hoc analysis can be found in Appendix B3.

4.1.5 *Discussion*

4.1.5.1 COMPOSITIONALITY–STRUCTURE HYPOTHESIS

Compositionality interacted significantly with structure: we found a larger acceptability difference between the idioms that we categorized as non-compositional and those that we categorized as compositional in all tested marked syntactic structures (prefield, LD, scrambling, and anaphor) than in the canonical word order baseline. While all tested structures differed significantly from the baseline in this respect, we failed to find differences between prefield, LD, scrambling, and anaphor in our post hoc analysis. We return to this finding and discuss what it means for the proposed hierarchies of syntactic flexibility in Section 4.6, after considering further structures in our Experiments 2–4. The observed compositionality–structure interaction is compatible with the view that syntactic flexibility is dependent on compositionality, i.e. compositional idioms are more flexible than non-compositional ones – provided that our categorization indeed reflects this property. We discuss this question and other potential sources of the effect in a by-item inspection of the data in Sections 4.5 and 4.6.

4.1.5.2 CONTEXT–STRUCTURE HYPOTHESIS

In line with Tabossi et al. (2009), we found a systematic effect of context on the acceptability of the marked syntactic structures. The polar question context facilitates all tested syntactic operations, as evidenced by the significant interaction

[13] In the cumulative link model analysis, a significant interaction was found for PREFIELD, but in the direction that compositional idioms were MORE acceptable (relative to non-idioms) in the PREFIELD condition than in the CANONICAL condition.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

between structure and context. However, since even the ratings for the anaphor condition were raised, for which no effect of information-structural factors was expected, the results do not inform us which of the marked syntactic structures require or prefer the affected constituent to be topical or contrastive – there seems to be a more general acceptability raising effect at play. It also unselectively affected non-idioms, compositional idioms, and non-compositional idioms alike (no three-way interaction between context, structure, and category). The results thus support the view that context is an important factor to systematically control in acceptability experiments on idiom flexibility, because it does have an effect on acceptability. But our context manipulation failed to provide further information to the question whether the tested structures are semantically restricted.

4.2 *Experiment 2 (German, replication of Experiment 1)*

One of the goals of the second experiment is to replicate Experiment 1 in order to corroborate our findings about the tested structures and to make sure our methodology yields replicable results. A second goal is to include a new set of additional (filler) stimuli for a better grasp of the acceptability level that we observe; these are discussed in Appendix A.

4.2.1 *Participants and procedure*

The recruitment procedure and experimental set-up were the same as for Experiment 1: 40 native speakers (different participants than in Experiment 1 but from the same population of mostly bachelor's degree students) completed the web-based questionnaire. In sum, 110 stimuli were presented to each participant (the same 90 critical items as in Experiment 1 and a new set of additional stimuli).

4.2.2 *Design and materials*

As for the critical items, the same design and materials were used as for Experiment 1. As in Experiment 1, additional stimuli besides the critical items were included. They were different from the ones in Experiment 1, but they again served to check assumptions about the participants' reactions to idiomatic expressions; see Appendix A.

4.2.3 *Results*

The results of Experiment 2 are summarized in Figure 2 and Table 3.

The same analysis procedure as reported above for Experiment 1 was used.

We found a significant two-way interaction between CONTEXT and STRUCTURE for each level of structure in the following direction: the acceptability difference between the canonical baseline and the other structures was reduced in the polarity

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

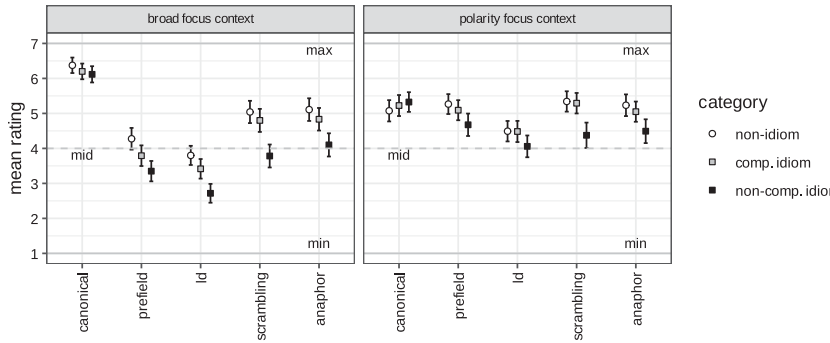

Figure 2
Experiment 2 – mean ratings (close-ended 1–7 point scale; error bars represent 95% confidence intervals).

Context	structure	Non-idioms	Comp. idioms	Non-comp. idioms
broad focus	canonical	6.38 (1.24)	6.20 (1.25)	6.12 (1.30)
broad focus	prefield	4.28 (1.73)	3.79 (1.65)	3.35 (1.63)
broad focus	LD	3.80 (1.53)	3.42 (1.56)	2.72 (1.50)
broad focus	scrambling	5.04 (1.78)	4.80 (1.84)	3.78 (1.83)
broad focus	anaphor	5.11 (1.81)	4.83 (1.80)	4.10 (1.85)
polarity focus	canonical	5.08 (1.70)	5.22 (1.68)	5.32 (1.57)
polarity focus	prefield	5.27 (1.59)	5.09 (1.60)	4.68 (1.79)
polarity focus	LD	4.49 (1.63)	4.48 (1.69)	4.06 (1.74)
polarity focus	scrambling	5.34 (1.63)	5.29 (1.64)	4.38 (2.03)
polarity focus	anaphor	5.23 (1.73)	5.05 (1.61)	4.49 (1.90)

Table 3
Results of Experiment 2 – mean ratings (standard deviations).

context in comparison to the broad focus context for prefield ($t = -8.04$, $p < 0.001$), LD ($t = -6.47$, $p < 0.001$), scrambling ($t = -5.79$, $p < 0.001$), and anaphor ($t = -3.30$, $p = 0.002$). We also found a significant two-way interaction between COMPOSITIONALITY and STRUCTURE for each level of structure when comparing compositional idiom to non-compositional idiom: the acceptability difference between these two idiom categories was larger than in the canonical baseline for prefield ($t = -2.27$, $p = 0.03$), LD ($t = -3.64$, $p < 0.001$), scrambling ($t = -3.02$, $p = 0.006$), and anaphor ($t = -4.18$, $p < 0.001$). We did not find significant interactions between COMPOSITIONALITY and STRUCTURE when comparing non-idiomatic to compositional idiom. We did not find significant three-way interactions. The complete fixed effect results can be found in Appendix B2.

Again, we conducted a post hoc analysis to see if prefield, LD, scrambling, and anaphor differed from each other with respect to the interaction with

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

COMPOSITIONALITY. As in Experiment 1, none of the pairwise comparisons was significant. The full results are shown in Appendix B3.

4.2.4 *Discussion*

The pattern that we see for the critical items is very similar to Experiment 1: again, we see an overall effect of context in that the ratings are higher in the polarity focus context, and again, we see a significant gap between non-compositional idioms and compositional ones in marked structures (toward lower ratings for non-compositional idioms), whereas non-idioms and compositional idioms behave similarly. This corroborates the robustness of the main findings of Experiment 1 and lends further support to the compositionality–structure and context–structure hypotheses. The replication of the context effect also alleviates a potential concern about Experiment 1: since context was a between-subject factor, the observed difference could have been due to unrelated differences between the two subject groups. This interpretation is much less likely in view of the replication of the pattern in Experiment 2.

An anonymous reviewer has correctly pointed out that the ratings for the structure with canonical word order are not exactly identical in Experiments 1 and 2 and in the two types of context and that it might therefore not be warranted to use it as a baseline. However, we think that having the baseline is crucial for our research questions. In our view, any slight differences between non-idioms, compositional idioms, and non-compositional idioms that we see in the canonical condition cannot have to do with syntactic flexibility (which cannot play a role in the canonical condition) but must be caused by independent, non-syntactic factors that influence the acceptability of our item groups. The same non-syntactic factors are likely to also be present in the marked structures. We thus think it is important to factor these out by not looking at each structure in isolation but by comparing the contrasts observed in the marked structures to those observed in the canonical condition; see also Appendix B. For this reason, we have decided to interpret the contrasts observed in the marked structures against the canonical baseline despite the potential caveat pointed out by the reviewer.

A further observation pointed out by a reviewer concerns the ratings in the polarity context, which are overall lower than in Experiment 1; we currently have no explanation for this.

4.3 *Experiment 3 (German, second set of structures)*

Experiments 1 and 2 showed a systematic difference between the idioms that we categorized as non-compositional and those that we categorized as compositional, but as for the tested structures, no consistent differences in the idiom behavior were found. The goal of Experiment 3 is to test further structures, namely passive, nominalization, and *which*-questions. For *which*-questions, there is the clear expectation that they should be restricted in their compatibility with idioms because *which*-questions are uncontroversially semantically restricted, as discussed in

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

Section 2. Including this condition thus should help to see whether our method is sensitive enough to detect structures that are more semantically restricted than others at all and what kind of pattern we should expect to see for them.

4.3.1 *Participants and procedure*

The recruitment procedure and experimental set-up were the same as for Experiments 1 and 2: 20 native speakers completed the online questionnaire; 130 stimuli were presented to each participant (108 critical items, the same 20 non-critical stimuli as in Experiment 2 and two additional ones; see Appendix A).

4.3.2 *Design and materials*

For this experiment, we dropped the context manipulation. In Experiments 1 and 2, we were mainly interested in testing structures that potentially involve an information-structural restriction, which motivated comparing two information-structurally different contexts, whereas this is not the focus of Experiment 3. Experiments 1 and 2 showed that the polarity context had a general facilitating effect: it raised the acceptability of marked structures in comparison to the broad focus context (without erasing crucial contrasts between conditions). In our view, it is beneficial if the items are as acceptable as possible with non-idiomatic idioms – this helps ensure that any observed deviations in acceptability can specifically be attributed to the manipulated factors (COMPOSITIONALITY and STRUCTURE) rather than to the lack of a suitable context. We thus chose to adopt the polarity focus rather than the broad focus context for Experiment 3.

As in Experiments 1 and 2, we tested three categories of VP (non-idioms, compositional idioms, and non-compositional idioms). The canonical baseline, prefield fronting, and LD were replicated from Experiments 1 and 2. In addition, passive, nominalization,[14] and *which*-questions were tested. The additional conditions are illustrated in (18). The same VPs as in Experiments 1 and 2 were used. For the *which*-questions, the contexts were slightly adjusted: instead of providing a yes or no question in the context (which would be odd to follow-up by another question), we only provided a declarative sentence.

(18) a. passive: *'Mary and Peter always used to fight against the unfair treatment of the trainees. Have they given up?'*

Nein,	das	Handtuch	wurde	bestimmt	nicht	geworfen!
no	the	towel	was	definitely	not	thrown

'No, the towel has definitely not been thrown!'

[14] A reviewer pointed out that the genitive construction we used in the nominalization condition might be stylistically at odds with some of the idioms (formal vs. informal register) and suggested using compounding as a more natural alternative (*Handtuchwerfen*, 'towel-throwing'). According to our intuition, the alternative construction would not fully work with all of our items (e.g. *? Zeitlichesegnen*), but we agree with the caveat about the register.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

b. nominalization: *'Mary and Peter always used to fight against the unfair treatment of the trainees. Have they given up?'*
Nein, den beiden ist bestimmt nicht zum Werfen des
no the two is definitely not to.the throwing of.the
Handtuchs zu Mute.
towel to spirit
'No, the two of them were definitely not in the mood for throwing in the towel.'

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

4.3.3 *Results*

The results of Experiment 3 are summarized in Figure 3 and Table 4.

The same analysis procedure as reported above for Experiments 1 and 2 was used.

We found a significant interaction between COMPOSITIONALITY and STRUCTURE for the following levels of structure when comparing compositional idiom to non-

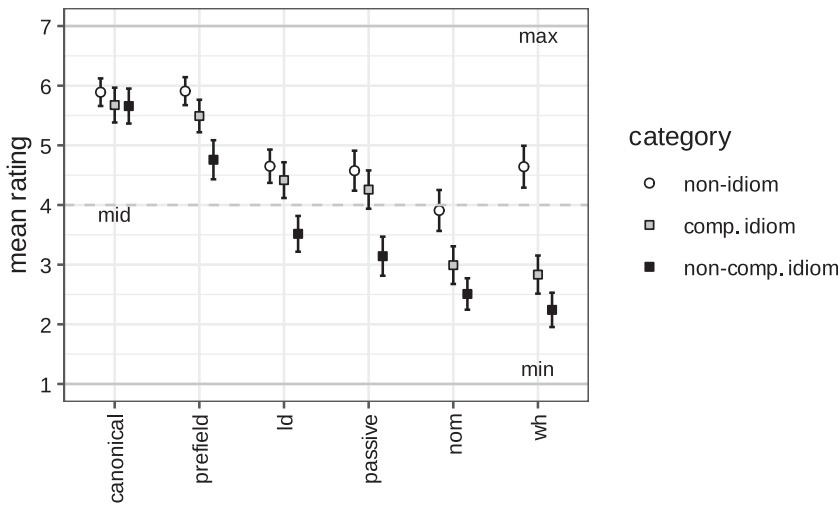

Figure 3
Experiment 3 – mean ratings (close-ended 1–7 point scale; error bars represent 95% confidence intervals).

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

Context	structure	Non-idioms	Comp. idioms	Non-comp. idioms
polarity focus	canonical	5.89 (1.29)	5.68 (1.63)	5.66 (1.64)
polarity focus	prefield	5.91 (1.31)	5.49 (1.52)	4.76 (1.83)
polarity focus	LD	4.65 (1.56)	4.42 (1.67)	3.52 (1.68)
polarity focus	passive	4.58 (1.87)	4.26 (1.79)	3.14 (1.83)
polarity focus	nominalization	3.91 (1.92)	2.99 (1.77)	2.51 (1.47)
polarity focus	which-question	4.64 (1.96)	2.83 (1.78)	2.24 (1.61)

Table 4
Results of Experiment 3 – mean ratings (standard deviations).

compositional idiom: the acceptability difference between these two idiom categories was larger than in the canonical baseline for prefield ($t = –2.56$, $p = 0.02$), LD ($t = –3.62$, $p < 0.001$), and passive ($t = –2.22$, $p = 0.04$). The difference between these idiom categories was not significantly different from the canonical baseline for nominalization ($t = –1.06$, $p = 0.30$) and *which*-question ($t = –1.31$, $p = 0.20$).[15] The interaction between COMPOSITIONALITY and STRUCTURE when comparing non-idiomatic to compositional idiom was only significant for *which*-question ($t = –3.62$, $p = 0.002$). The difference between these idiom categories was not significantly different from the canonical baseline for prefield ($t = –0.74$, $p = 0.47$), LD ($t = –0.07$, $p = 0.95$), passive ($t = –0.20$, $p = 0.84$), and nominalization ($t = –1.57$, $p = 0.13$).[16] The complete fixed effects model results can be found in Appendix B2.

In post hoc comparisons, we tested whether prefield, LD, passive, nominalization, and *which*-question differed from each other with respect to the interaction with COMPOSITIONALITY. The numerically largest contrasts were found when comparing the *which*-question to all other structures with respect to the gap between non-idioms and compositional idioms; the contrast was significant for *which*-question vs. LD. See Appendix B3 for detailed results.

4.3.4 *Discussion*

For prefield fronting and LD, the results confirm our previous findings from Experiments 1 and 2: there is a statistically significant gap between compositional idioms and non-compositional ones, while we failed to find a significant gap

[15] In the cumulative link model analysis, the interaction was significant for *which*-question. Deviations between the linear model (which treats the response value as numerical) and cumulative link model (treating the response variable as ordinal) can occur, e.g. when two conditions do not differ much in their means, but the distribution of individual rating categories (how frequently '1' / '2' / '3' /… was chosen) is different. Our approach to discrepancies like this is to refrain from basing conclusions on findings that are not consistently confirmed by all statistical analyses that we ran. This particular deviation does not affect any of the conclusions drawn in Section 4.3.4.

[16] In the cumulative link model analysis, the interaction was significant for NOMINALIZATION. Again, this does not affect the conclusions drawn in Section 4.3.4.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

between compositional idioms and non-idioms. The newly tested passive structure also conforms to this pattern. Nominalization deviates from the other structures in that compositional and non-compositional idioms do not differ significantly; rather, there is a trend toward a larger difference between non-idioms and compositional idioms. For *which*-questions, this difference between non-idioms and compositional idioms is significant and numerically large. This shows that our method is sensitive enough to reflect that not all syntactic structures behave alike with idioms – *which*-questions, which we assume to definitely be semantically restricted (as discussed in Section 2), are degraded with a larger part of the tested idioms, including at least some of those that we categorized as compositional intuitively. This in turn suggests that if the other tested structures are semantically restricted, this restriction is weaker. The finding for *which*-questions also indicates that the compositional vs. non-compositional idioms comparison is not the only relevant one: semantic restrictiveness can also be reflected in a larger difference between non-idioms and compositional idioms. The results of Experiment 3 thus underline the importance of including non-idioms in the experiment for comparison – without this baseline, we would see that some marked structures are less acceptable than canonical word order, but we would not be able to assess how much of the acceptability decrease is specific to idioms.

It is interesting to note that nominalization shows a trend in the same direction as the *which*-question condition. The robustness of this trend and the reason behind it is something that would be worth exploring in future research. Following a reviewer's suggestion, we think referentiality (of the DP) is a semantic property that would be worth exploring further in this context.

4.4 *Experiment 4 (English)*

In the fourth and final experiment reported here, we test a range of English structures: nominalization with and without 'of', passive, a cleftlike structure, and pronominalization. These structures have been experimentally investigated before using the similarity-rating method by Gibbs & Gonzales (1985) and Gibbs & Nayak (1989). The first goal of retesting is to see if we can find further corroboration of the assumption that semantic restrictiveness of syntactic structures is detectable in our paradigm: among the English structures that we test is a cleftlike structure for which we assume that it is clearly semantically restricted and which can thus serve as a further test case, similar to the *which*-questions tested in Experiment 3: it should show a clearly different pattern than the canonical baseline. If this premise holds, a second goal is to get a clearer impression of the more controversial conditions: our methodology allows for a more direct comparison of the syntactic structures than the previous similarity-rating studies that were designed with different primary research goals. In particular, it will be informative to see whether the passive is indeed limited with respect to idiomatic expressions (as claimed, among others, by Nunberg et al. 1994) and whether we

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

can find support for a difference between nominalization with and without 'of', e.g. 'Mary's kicking (of) the bucket', as claimed by Fraser (1970) and not consistently resolved by previous studies.[17]

4.4.1 *Participants and procedure*

A total of 20 native speakers of English were recruited at University College London (Division of Psychology and Language Sciences, via Sona Systems) and at Trinity Hall, Cambridge. They were paid for participation. The questionnaire was set up in the same way as described for Experiments 1–3. Each participant saw 128 stimuli (108 critical items and 20 additional stimuli; the latter corresponded to the ones used in Experiments 2 and 3).

4.4.2 *Design and materials*

As in Experiment 3, we only tested one type of context (polarity focus). We selected the idioms for our study based on the results of Gibbs & Nayak's (1989) categorization task: we adopted six of the idioms that were mostly categorized as 'normally decomposable' by their participants (we assume this to be the category most closely corresponding to our 'compositional' idioms in Experiments 1–3) and five of the idioms that were mostly categorized as 'non-decomposable'. To the latter group, we added one idiom (bite the bullet) that we considered to be non-compositional and that was not in Gibbs & Nayak's list, based on the intuition that the other reported idioms might not be familiar enough to the young speakers of British English whom we intended to test. The idioms are shown in Table 5.

We added six non-idiomatic VPs: *forget the timer*, *eat the cake*, *cut down the hedges*, *reveal the trick*, *paint the car*, and *throw away the cutlery.*

We tested six syntactic structures, as illustrated in (19). We refer to condition (19d) as 'cleftlike', because it shares properties with clefts: it is a biclausal structure involving a copular construction and relativization. We chose this structure rather than it-clefts ('It is the question that he popped') or pseudo-clefts ('What he would never pop is the question') because it was easier to construct plausible items while keeping the context constant. It-clefts are semantically restricted because they have exhaustive semantics roughly along the lines of 'There is only one x for which it is true that Harry would never pop x, and x is the question' (see e.g. Velleman et al. 2012 for a formal analysis). Our cleftlike condition does not involve exhaustivity, but its meaning can be paraphrased as 'There is a set S of elements that Harry would never pop, and the question is an element of S' – similar to our reasoning concerning

[17] As noted in the discussion of Experiment 3, with respect to nominalization, our goal is limited to testing the empirical claim that the two nominalization structures show very different degrees of restrictiveness. Discussing potential reasons for the restrictiveness, for which semantic explanations as in the case of *which*-questions or clefts are less straightforward, is beyond the scope of this paper.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

Category	Idiom	Paraphrase of the figurative meaning
non-comp.	kick the bucket	'to die'
non-comp.	raise the roof	'to complain loudly and angrily'
non-comp.	shoot the breeze	'to chat aimlessly'
non-comp.	chew the fat	'to chat or gossip'
non-comp.	play the field	'to date a variety of people'
non-comp.	bite the bullet	'to party excessively'
compositional	pop the question	'to propose to somebody'
compositional	lay down the law	'to give strict orders'
compositional	break the ice	'to reduce the social tension'
compositional	miss the boat	'to lose an opportunity'
compositional	hit the sauce	'to drink alcohol'
compositional	clear the air	'to reduce the social tension'

Table 5
English idioms used in Experiment 4.

the semantics of *which*-questions and *it*-clefts, this should also be impossible to interpret or evaluate if the expression *the question* lacks an individual denotation.

(19) 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. nominalization with 'of': Yes, he did, but I don't really want to talk about Harry's popping of the question at the moment.
   c. passive: Of course not, the question would definitely never be popped by such an incorrigible player!
   d. cleftlike: Of course not, the question is something that he would definitely never pop!
   e. anaphor: I'd say so… even though no one thought he would ever pop the question, he obviously did pop it.
   f. nominalization without 'of': Yes, he did, but I don't really want to talk about Harry's popping the question at the moment.

4.4.3 *Results*

The results of Experiment 4 are summarized in Figure 4 and Table 6.

The same analysis procedure as for Experiments 1–3 was used.

We did not find a significant interaction between COMPOSITIONALITY and STRUCTURE when comparing compositional idiom to non-compositional idiom: the acceptability difference between these two idiom categories was not different from the canonical baseline for any of the tested structures (nominalization with 'of': $t = 0.52$, $p = 0.60$; passive: $t = -1.62$, $p = 0.12$; cleftlike: $t = 0.45$, $p = 0.66$; anaphor: $t = -0.82$, $p = 0.42$; and nominalization without 'of': $t = 1.55$, $p = 0.13$). The interaction between COMPOSITIONALITY and STRUCTURE when comparing

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

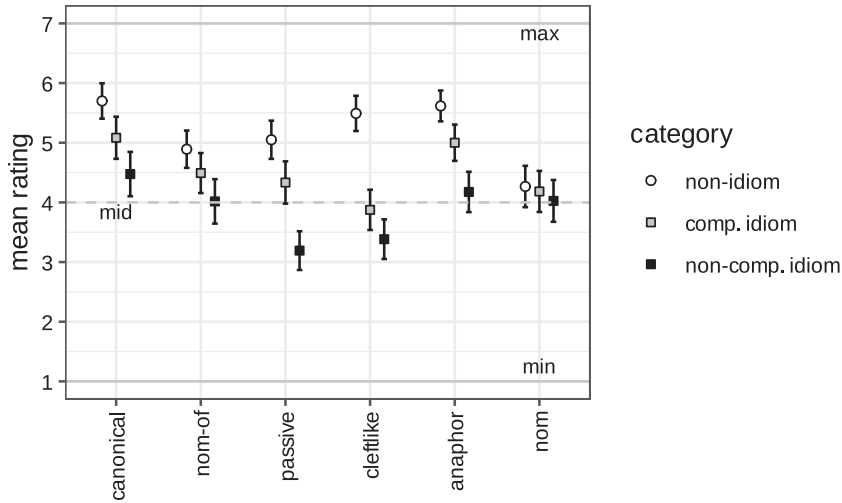

Figure 4
Experiment 4 – mean ratings (close-ended 1–7 point scale; error bars represent 95% confidence intervals).

Context	structure	Non-idioms	Comp. idioms	Non-comp. idioms
polarity focus	canonical	5.70 (1.66)	5.08 (1.97)	4.47 (2.08)
polarity focus	nominalization with 'of'	4.89 (1.75)	4.49 (1.87)	4.02 (2.08)
polarity focus	passive	5.05 (1.79)	4.33 (1.98)	3.19 (1.82)
polarity focus	cleftlike	5.49 (1.65)	3.88 (1.89)	3.38 (1.86)
polarity focus	anaphor	5.62 (1.44)	5.00 (1.70)	4.18 (1.89)
polarity focus	nominalization without 'of'	4.27 (1.94)	4.18 (1.92)	4.03 (1.96)

Table 6
Results of Experiment 4 – mean ratings (standard deviations).

non-idiomatic to compositional idiom was only significant for cleftlike ($t = –3.89$, $p = < 0.001$). The difference between these idiom categories was not significantly different from the canonical baseline for nominalization with 'of' ($t = 0.85$, $p = 0.40$), passive ($t = –0.30$, $p = 0.76$), anaphor ($t = 0.00$, $p = 1$), and nominalization ($t = 1.88$, $p = 0.07$). The complete model results for the fixed effects can be found in Appendix B2.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

In post hoc comparisons, we tested whether the structures differed from each other with respect to the interaction with COMPOSITIONALITY. The numerically largest contrasts were found when comparing the cleftlike condition to all other structures with respect to the gap between non-idioms and compositional idioms; the contrast was significant for cleftlike vs. nominalization with 'of', cleftlike vs. nominalization without 'of', and cleftlike vs. anaphor. See Appendix B3 for detailed results.

4.4.4 *Discussion*

In comparison to Experiments 1–3, a notable difference is that the compositionality categories are further apart already in the canonical word order, which might e.g. be due to the participants being familiar to different degrees with the individual idioms. This stresses the importance of including this baseline: without it, the patterns in the other conditions would be prone to misinterpretation. For example, even though we see clear contrasts between the categories in the passive, these are not significantly larger than in the canonical baseline, indicating that this is not a structure-specific effect. Nevertheless, for the cleftlike structure, we see a clear deviation from the baseline and from all other conditions, in line with the expectations. As for the nominalization structures, they do not show larger differences between the levels of compositionality than the canonical baseline (there is even a trend in the opposite direction). This strongly suggests that they are not particularly restricted with respect to idioms and that they are not at opposite ends of a hierarchy in this respect, contra Fraser's (1970) proposal.[18]

4.5 *Exploratory by-item analysis*

Our conclusions about the effect of compositionality on idiom flexibility rely on the assumption that our categorization of 'compositional' and 'non-compositional' idioms really reflects the intended property. We conducted a post hoc by-item analysis to explore other potential influencing factors and individual differences between the tested idioms.

For this, we pooled the data from all German experiments. The means for canonical word order, prefield, and LD are therefore based on data from Experiments 1–3; the means for scrambling and anaphor are based on data from Experiments 1 and 2; and the means for passive, nominalization, and *which*-question are based on data from Experiment 3. The results for each of the 18 items (12 idioms and six non-idioms) are shown in Figure 5. The by-item results for the English

[18] A reviewer asks what could be the reason for the lower acceptability of German nominalization in Experiment 3 in comparison to the English structures in Experiment 4. This could be related to the potential problem with deviating register of our German nominalization condition in footnote 13.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

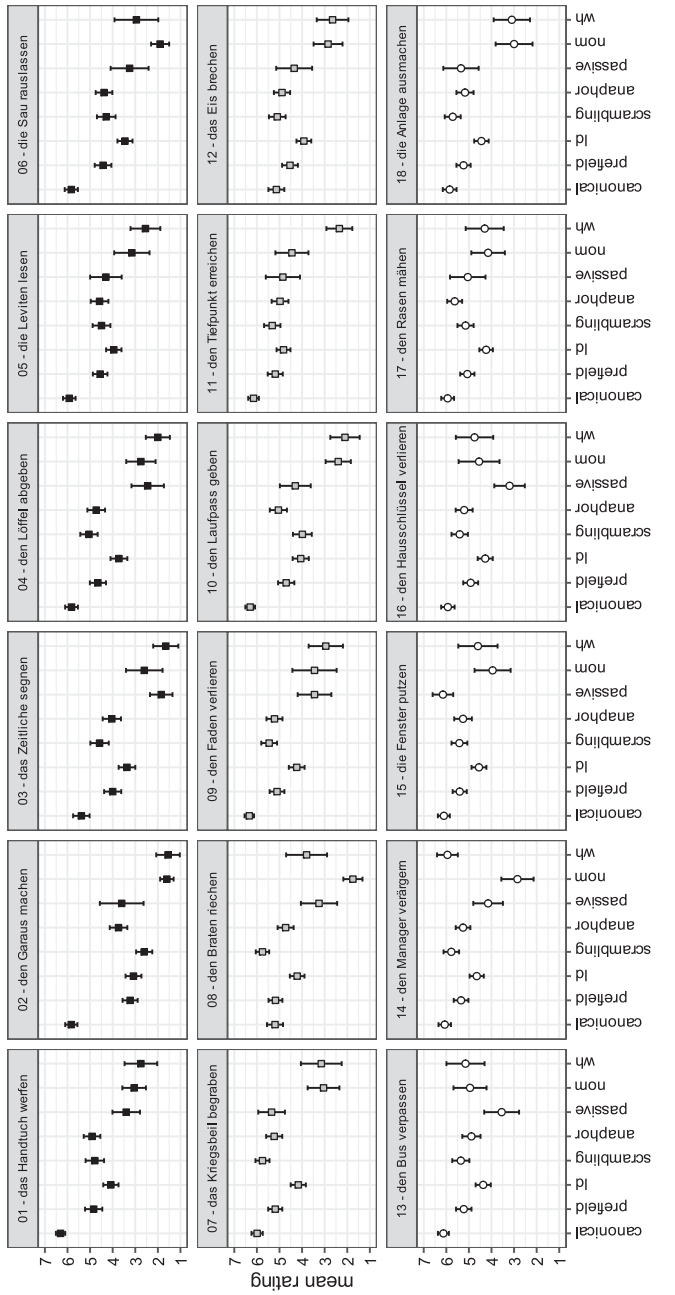

Figure 5
By-item results based on pooled data from Experiments 1–3.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

idioms can be found in the OSF repository. By-participant results are presented and discussed in Appendix C.

Visual inspection shows that while on average, the non-compositional idioms (in the first row of Figure 5) show a larger acceptability difference between canonical word order and the other structures than the compositional ones (second row), there are also individual differences. When we focus e.g. on the distance between canonical and prefield, the non-compositional idioms that show the largest acceptability gap involve a unical element, i.e. one that does practically not occur outside of the idiom, namely 'GARAUS'. It is notable that the compositional idiom that shows the largest gap is 'den LAUFPASS geben', which also involves a compound that only occurs within this idiom. Another property that these two VPs have in common is that they require a dative object; this also applies to 'die LEVITEN lesen'.

A further observation is that for the passive condition, there are large individual differences, even within our compositionality categories: for example, within the idioms categorized as compositional, there are some that can be felicitously passivized (*das Kriegsbeil begraben*) and some that cannot (d*en Faden verlieren*).

4.6 *General discussion*

Our results are compatible with the view that there are syntactic structures that are semantically restricted, meaning that some types of constituents (in particular, expressions that do not have an individual meaning, like elements that are part of an idiom) cannot appear in them felicitously. For German prefield, LD, scrambling, anaphor, and passive, we found that their acceptability with a part of our idiom set (the idioms categorized as non-compositional) is significantly lowered in comparison to non-idioms; for German *which*-questions and the English cleftlike structure, we found lowered acceptability with the idioms categorized as compositional in comparison to non-idioms.

Our results are informative with respect to proposed restrictiveness hierarchy of the tested syntactic structures. For English, we observed similar patterns for nominalization with and without 'of', which speaks against the assumption that these two structures are at opposite ends of a restrictiveness hierarchy with respect to idioms, as proposed by Fraser (1970). As for German, our results go against the view that LD and scrambling are more restricted than *which*-questions while prefield fronting is almost unrestricted, as proposed by G. Müller (2000, 2019). Our results do not exclude the possibility that there are more fine-grained differences between prefield fronting, LD, and scrambling that we failed to detect, but if they exist, they seem to be perceived as more subtle by most speakers than has been proposed. Regarding the German passive, the high variability observed in the by-item analysis suggests that compositionality is not the (only) influencing factor here – the availability of passivization seems to depend not only on compositionality but also on further properties that varied both between and within our idiom categories.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

Our experiments also provide methodological insights. First, context plays a role: there was a consistent increase of acceptability in the polarity context in comparison to the broad focus context in Experiments 1 and 2, suggesting that this is a factor that should not be neglected when estimating the syntactic flexibility of idioms; without a suitable context, it might be underestimated how acceptable even non-compositional idioms can be in non-canonical structures. Second, as shown in particular by Experiment 4 on English, it is important to include non-idioms and sentences with canonical word order as a baseline; otherwise, the importance of observed contrasts in marked syntactic structures could be overestimated.

The next logical question is what the findings mean for grammatical models of the tested structures. To answer it, it is helpful to take into account absolute ratings and effect sizes.[19] An observation that we can make based on visual inspection of our data is that there are some structures which are consistently rated above the midpoint of our scale in the facilitating polarity context (4 on the 7-point scale), even with non-compositional idioms (in particular, the German prefield condition), while others are clearly rated in the range below the midpoint when occurring with idioms (e.g. *which*-questions). This observation suggests that prefield fronting is semantically less restricted than *which*-questions; but does it support models that predict parts of non-compositional idioms to be ungrammatical in *which*-questions and grammatical in the prefield position? To evaluate this, we need to compare the observed effect sizes to other types of grammatical violations in future work.

To illustrate this suggestion, take for example Frey's (2010) theory of the German prefield. In this model, an object occurring in the prefield position needs to have a certain information-structural interpretation (emphasis or contrast), which is not possible if it is a part of a non-compositional idiom. However, if this requirement is not met, the model does not predict ungrammaticality. The requirement is encoded in terms of a conventional implicature that is violated when the constituent in the prefield does not have the required interpretation. In order to test the model's exact predictions empirically, we could compare the effect size observed for test sentences with idioms (as in our experiments) to other well-established cases of conventional implicature violations (Potts 2007), ideally within one experimental set-up. Our experiments provide a starting point for such more in-depth investigations.

4.7 *Outlook*

Our experiments open up further directions for future research. First, potential differences between the syntactic structures could be investigated in more detail in an experiment specifically designed toward detecting more fine-grained contrasts between a smaller number of structures.

[19] We thank a reviewer for pointing out the importance of absolute values and effect sizes.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

Second, our manipulation of the factor compositionality in the German experiments was based on our own intuitive categorization. There are certainly cases that other speakers perceive differently. For example, as pointed out by a reviewer, while we categorized *das Kriegsbeil begraben* 'to bury the hatchet' as compositional, based on the intuition that *the hatchet* figuratively refers to a conflict and *burying* means ending it, a different plausible view is to see the action as a whole as a symbol for making peace, similar to how throwing in the towel (which we categorized as non-compositional) in the boxing ring symbolizes surrendering. To take into account the subjectiveness of compositionality and to increase generalizability, compositionality ratings could be collected from a large number of speakers in future experiments (following Hubers et al.'s 2019 assessment that judgments of idiom properties can be collected reliably). This would also make a more fine-grained approach to compositionality possible, rather than splitting the idioms into two categories.

Third, as it has been shown that compositionality correlates with other idiom properties, there is the question of whether there really is a causal relation between this factor and syntactic flexibility. The by-item analysis (Section 4.5) suggested that unicality is a factor in which the tested idioms vary and which could be related to limited syntactic flexibility. If this is the case, then a part of the observed difference between the idiom categories could be attributed to the fact that our set of non-compositional idioms contained several unical elements – *Garaus*, *das Zeitliche* (nominalized form of *zeitlich* 'temporal'), and *Leviten.* Unicality can be seen as an extreme case of non-compositionality; it is thus important to check whether our conclusions hold up when a broader spectrum of idioms is tested. More generally, it could play a role for syntactic flexibility in how likely the DP is to occur with this verb and vice versa (conditional frequencies, as discussed by Müller & Englisch 2020) or whether DP and verb independently co-occur with similar words and phrases (similarity of semantic vectors, as discussed by Gehrke & McNally 2019). A related question is what role unicality plays in the categorization of an idiom as compositional or non-compositional. A reviewer expressed the view that idioms consisting of a unical element and a semantically light verb like *den Garaus machen* 'to kill' are special in that neither the DP nor the verb has a figurative meaning: since the DP only occurs in the idiomatic sense, there is no distinction between literal and idiomatic meaning; and since the verb is semantically light, its function within the idiom is not really different from literal uses, either. In our experiments, we decided to subsume this case under 'non-compositional', following the criterion that it is not possible to assign individual figurative meanings to the DP and verb.[20] However, we agree that this reinforces the view that idioms are a

[20] The reviewer suggested that the same reasoning applies to *den Laufpass geben* 'to break up', which we categorized as compositional. Our intuition was different here than for *den Garaus machen* 'to kill', because we perceive both *geben* (literal meaning: 'to give') and *Laufpass* (a compound consisting of *Lauf* 'run' and *Pass* 'pass', which do occur individually outside of the idiom) as having more semantic content and thus more potential for figuration or metaphor.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

highly heterogeneous set of expressions and that a more fine-grained analysis is desirable in follow-up studies.

Fourth, our findings are based on 12 idioms that were deliberately homogeneous with respect to their structure (verb and definite direct object, without negation) to avoid confounds and to ensure that all tested syntactic structures could be constructed with them. It is desirable to test whether our findings generalize to other types of idioms. On the other hand, there are also structural properties in which our items varied (e.g. whether they require a dative object), which could be controlled and studied more systematically.

A fifth and final direction for future research (suggested by a reviewer) is embedding idiomatic materials like ours in acceptability studies with a higher proportion and wider range of non-idiomatic sentences, as figurative expressions are much rarer in natural speech than in our experiments.

5. Conclusion

In a series of four acceptability rating experiments, we have investigated the syntactic flexibility of idioms, varying the factors compositionality, context, and syntactic structure. Let us return to the research questions from Section 1: (i) Can semantic restrictiveness explain to what extent syntactic structures are compatible with idioms, and (ii) Can compositionality explain the differences in the flexibility of idioms?

With respect to question (ii), our results are compatible with the view that syntactic flexibility depends on compositionality (in line with the findings by Gibbs & Nayak 1989 for English): while we did not find a significant difference in flexibility between compositional idioms and non-idioms for most of the tested structures, non-compositional ones were degraded in non-canonical structures quite consistently. A caveat based on a by-item analysis is that other factors correlating with our compositionality categorization could contribute to the effect, in particular the presence of elements that uniquely occur within the idiom.

As for question (i), our results provide evidence that our method is able to detect syntactic structures that are incompatible with a larger set of idioms than others. In particular, German *which*-questions and English cleftlike structures were found to be generally less acceptable with idioms (even compositional ones) than with non-idioms, while most of the other tested structures were only significantly degraded with non-compositional idioms. This is compatible with the view that *which*-questions and cleftlike structure have a higher degree of semantic restrictiveness than the other structures. For other structures that have been proposed to differ in this respect (e.g. German scrambling vs. prefield fronting or LD), we did not find consistent contrasts.

Competing interests

The authors declare none.

https://doi.org/10.1017/S0022226723000105 Published online by Cambridge University Press

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S0022226723000105.

REFERENCES

Ackerman, Farrell & Gert Webelhuth. 1993. Topicalization and German complex predicates. Ms., University of California, San Diego/ University of North Carolina.

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

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

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015b. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 1(67), 1–48.

Bresnan, Joan. 1982. The passive in lexical theory. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 3–86. Cambridge, MA: MIT Press.

Cardinaletti, Anna. 1986. Topicalization in German: Movement to Comp or base-generation in Top? *GAGL (Groninger Arbeiten zur germanistischen Linguistik)* 28, 202–231.

Christensen, Rune Haubo Bojesen. 2019. ordinal: Regression models for ordinal data. R package version 2019.12–10. https://CRAN.R-project.org/package=ordinal (accessed 10 March 2023).

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

Fanselow, Gisbert. 2004. Cyclic phonology-syntax interaction: Movement to first position in German. *Universitätsverlag Potsdam Working Papers of the SFB 632: Interdisciplinary studies on information structure* 1, 1–42.

Fanselow, Gisbert. 2010. Scrambling as formal movement. In Ivona Kučerová & Ad Neeleman (eds.), *Contrasts and positions in information structure*, 267–295. Cambridge: Cambridge University Press.

Fanselow, Gisbert & Denisa Lenertová. 2011. Left peripheral focus: Mismatches between syntax and information structure. *NLLT* 29, 169–209.

Fanselow, Gisbert. 2018. Zur Flexibilität von Idiomen im Deutschen. *Colloquia Germanica Stetinensia* 27, 115–134.

Fellbaum, Christiane. 2019. How flexible are idioms? A corpus-based study. *Linguistics* 57(4): 735–767.

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

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

Frey, Werner. 2004b. A medial topic position for German. *Linguistische Berichte* 199, 153–190.

Frey, Werner. 2004c. Notes on the syntax and the pragmatics of German left dislocation. In Horst Lohnstein & Susanne Trissler (eds.), *The syntax and semantics of the left periphery*. Berlin: De Gruyter Mouton.

Frey, Werner. 2010. Ā-movement and conventional implicatures: About the grammatical encoding of emphasis in German. *Lingua* 120:1416–1435.

Gehrke, Berit & Louise McNally. 2019. Idioms and the syntax/semantics interface of descriptive content vs. reference. *Linguistics* 57(4): 769–814.

Gibbs, Raymond W. & Gayle P. Gonzales. 1985. Syntactic frozenness in processing and remembering idioms. *Cognition* 20, 243–259.

Gibbs, Raymond W. & Nandini P. Nayak. 1989. Psycholinguistic studies on the syntactic behaviour of idioms. *Cognitive Psychology* 21, 100–138.

Goldberg, Adele E. 2006. Compositionality. In Nick Riemer (ed.), *The Routledge Handbook of semantics*. London/New York: Routledge.

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

Hubers, Ferdy, Catia Cucchiarini, Helmer Strik & Ton Dijkstra. 2019. Normative data of Dutch idiomatic expressions: Subjective judgments you can bank on. *Frontiers in Psychology* 10, 1075.

Jackendoff, Ray. 2008. Construction after construction and its theoretical challenges. *Language* 84, 8–28.
Jacobs, Joachim. 2001. The dimensions of topic–comment. *Linguistics* 39(4), 641–681.
Karttunen, Lauri (1977). Syntax and semantics of questions. *Linguistics and Philosophy* 1, 3–44
Katz, Jerrold J. 1966. *The philosophy of language*. New York: Harper and Row.
Kay, Paul & Ivan A. Sag. 2014. *A lexical theory of phrasal idioms*. Ms., University of California, Berkeley & Stanford University.
Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software* 13(82).
Leiner, Dominik. 2018. SoSci Survey [software]. http://www.soscisurvey.com (accessed 10 March 2023).
Maher, Zachary. 2013. *Opening a can of worms: idiom flexibility, decomposability, and the mental lexicon.* BA/MA thesis, Yale University.
Müller, Gereon. 2000. *Idioms and transformations.* Presented to the GGS meeting, Potsdam.
Müller, Gereon. 2019. *Syntactic strength: A new approach.* Presented to the (Computer-)Linguistisches Kolloquium, University of Stuttgart.
Müller, Stefan. 2010. Persian complex predicates and the limits of inheritance-based analyses. *Journal of Linguistics* 46(3), 601–655.
Müller, Gereon & Johannes Englisch. 2020. Extraction from NP, frequency, and minimalist gradient harmonic grammar. Presented to the DGfS meeting in Hamburg, AG Modeling Gradient Variability in Grammar.
Nunberg, Geoffrey. 1977. *The pragmatics of reference.* Graduate Center dissertation, City University of New York
Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3), 491–538.
Potts, Christopher. 2007. Conventional Implicatures: A Distinguished Class of Meanings. In Gillian Ramchand and Charles Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces*.
R Core Team. 2016. R: A language and environment for statistical computing. https://www.R-project.org.
Soehn, Jan-Philipp. 2006. *Über Bärendienste und erstaunte Bauklötze: Idiome ohne freie Lesart in der HPSG*. Frankfurt am Main: Peter Lang.
Tabossi, Patrizia, Rachele Fanari & Kinou Wolf. 2008. Processing idiomatic expressions: effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(2), 313–327.
Tabossi, Patrizia, Kinou Wolf & Sara Koterle. 2009. Idiom syntax: idiosyncratic or principled? *Journal of Memory and Language* 61, 77–96.
Velleman, Dan B., David Beaver, Emilie Destruel, Edgar Onea & Liz Coppock. 2012. It-clefts are IT (inquiry terminating) constructions. *Proceedings of SALT* 22, 441–460.
Wasow, Thomas, Geoffrey Nunberg & Ivan A. Sag. 1984. Idioms: An interim report. In Shiro Hattori and Kazuko Inoue (eds.), *Proceedings of the XIIIth International Congress of Linguists*, 102–15. Tokyo: Nippon Toshi Center.
Webelhuth, Gert & Farrell Ackerman. 1999. A lexical-functional analysis of predicate topicalization in German. *Journal of Germanic Linguistics* 11(1), 1–61.
Weinreich, Uriel. 1969. Problems in the analysis of idioms. In Jaan Puhvel (ed.), *Substance and structure of language*, 23–81. Berkeley: University of California Press.

Authors' addresses: *(Wierzba)*
Universität Potsdam, Department Linguistik, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany

(Brown)
Universität Potsdam, Department Linguistik, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany; Université de Lausanne, Section d'anglais, Quartier UNIL-Chamberonne, Bâtiment Anthropole 5119, 1015 Lausanne, Switzerland

(Fanselow)
Universität Potsdam, Department Linguistik, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany