Appendix to the paper "The syntactic flexibility of German and English idioms: evidence from acceptability rating experiments"

by Marta Wierzba, J.M.M. Brown, and Gisbert Fanselow

APPENDIX A: ADDITIONAL STIMULI

Besides the critical stimuli described in the main text, our experiments also included a number of additional (filler) materials, which we will describe and discuss in more detail here.

A.1 *Additional stimuli in Experiment 1*

In Experiment 1, the additional stimuli included: six stimuli with a made-up idiom, six stimuli with an idiom in which one of the words was replaced by different word, and six stimuli in which a verb was used in its literal and idiomatic sense within the same sentence with a pun-like effect. We included these stimuli types to get an impression of the acceptability of the syntactically manipulated idioms in comparison to other (non-syntactic) manipulations of idioms. Types 1 and 2 served to test how participants would react to complete or partial deviations from standard idioms at the lexical level (replacement of words), the expectation being that those should be strongly degraded. Type 3 served to address the following consideration: it is conceivable that participants might be more willing to tolerate a grammatical violation in their judgments when it comes to idioms because as a non-literal part of the language, idioms are often subject to word-play and puns. To get a first impression of the acceptability range of pun-like items and to see how they compare to the critical items, this additional filler type was included.

The types of additional stimuli included in Experiment 1 are illustrated in (1).

(1) (a) Type 1: made-up idiom

Ich habe gehört, dass Michi im Philosophieseminar einen Vortrag gehalten hat. Denkst du, er war gut? – Nein, das Telefon hat er bestimmt nicht zum Klingeln gebracht!

'I heard that Michi gave a presentation in the philosophy seminar. Do you think he did a good job? – No, he certainly did not make the telephone ring!' (no idiomatic meaning)

(b) Type 2: idiom with replaced word

Warum kommt Lisa eigentlich nicht mehr zum Fußballtraining? – Die Flinte hat sie aufs Dach geworfen.

'Why does Lisa not attend the soccer training anymore? – She threw the gun onto the roof.' (based on die Flinte ins Korn werfen, lit. 'to throw the gun into the grain' = 'to give up')

(c) Type 3: juxtaposition of literal and idiomatic reading (pun)

Wieso ist Mario so aufgebracht, hat er was verloren? – Er hat seinen Geldbeutel und den Kopf verloren.

'Why is Mario so upset, did he lose something? – He lost his wallet and his head.' (based on den den Kopf verlieren, lit. 'to lose one's head' = 'to lose one's self-control')

Analysis of the stimulus types showed a mean rating of 1.50 (standard deviation: 1.17) for the made-up idioms[1], 1.86 (standard deviation: 1.51) for the lexically altered idioms,

1 One of the six filler items with made-up idioms was excluded from analysis due to a typo.

and 3.68 (standard deviation: 1.98) for the stimuli combining a literal and idiomatic interpretation of the same word.

Comparison of the critical items of Experiment 1 to these additional materials shows that in the polarity focus context, the mean ratings for the critical items are above the means of non-critical stimulus types 1-2 across all conditions. This tentatively suggests that with a suitable context, all idioms – even non-compositional ones – have a certain degree of syntactic flexibility: fronting, left-dislocating, scrambling or pronominalizing a part of a non-compositional idioms does not make the sentence as unacceptable in this kind of task as including a clear lexical mistake or wordplay-like mixing of idiomatic and non-idiomatic readings.

A.2 *Additional stimuli in Experiments 2-4*

In Experiment 2, different additional stimuli were used than in Experiment 1, with the goal of testing further assumptions about the participants' reaction to idiomatic expressions. There were two types. The first type were idioms with a DP that usually occurs in the singular form within the idiom. We tested a singular and a plural version. The second type contained minimizers like *Schimmer* 'clue' (lit. 'shimmer'), which usually only occur under negation (*keinen Schimmer haben* 'have no clue', *#einen Schimmer haben* 'have a clue'). We thank Balázs Surányi for the idea to include minimizers for comparison. We tested them with and without negation. These stimulus types were also included in Experiments 3 and 4.

The reasoning behind these types of fillers was that they can help us to see how high the acceptability level of the critical items is in relation to further types of deviations from the standard form of idiomatic expressions. In contrast to the lexical alternations

that were included in the additional stimuli of Experiment 1, the singular/plural alternation tested here does not completely distort the meaning of the idiom, but it clearly deviates from the form in which the idiom normally occurs. All contexts were constructed in such a way that there is a potential pragmatic motivation for the deviation, i.e., the context always introduced multiple individuals. Similarly, the stimuli with minimizers were constructed in such a way that in principle both versions would make sense in the context, but only in one version, the formal requirement of occurring under negation is satisfied. This gives us an additional impression (based on non-structural manipulations) of how much the acceptability drops in this type of task in cases in which a deviation from the standard form of an idiom is pragmatically motivated.

The types of additional stimuli included in Experiment 2 are illustrated in (2-3).

(2) Filler type 1: number manipulation

Was haben die vier Studenten aus der hintersten Reihe denn gemacht, als die Professorin sie plötzlich an die Tafel nach vorne bat? *'What did the four students in the back row do when the professor asked them to come to the blackboard?'*

– Na, sie haben in den sauren Apfel / in die sauren Äpfel gebissen und sind nach vorne gekommen.

'Well, they swallowed the pill/the pills and came to the front.' (lit.: they bit into the sour apple / sour apples)

(3) Filler type 2: minimizers

Lars wollte dir doch beim Einrichten deines neuen Laptops helfen, wie ist das gelaufen? *'Lars wanted to help you with setting up your new laptop, right? How did it go?'*

– Er hat wirklich keinen Schimmer / einen Schimmer von Computern. *'He really has no clue / a clue about computers.'*

In Experiment 2, for the first group of fillers (number manipulation), a mean rating of 5.63 (standard deviation: 1.61) was found for the idioms with the normal singular form, and a mean rating of 3.92 (standard deviation: 2.14) for the version with a plural form. For the second group of fillers (minimizers), a mean rating of 6.00 was found for the version where the minimizer was licensed by negation, and 3.08 without negation.

In Experiment 3, for the filler group with the number manipulation, a mean rating of 5.65 (standard deviation: 1.75) was found in the singular condition and 3.30 (standard deviation: 2.04) in the plural condition. For the filler group with minimizers, a mean rating of 4.86 (standard deviation: 2.21) was found in the condition with negation and 3.57 (standard deviation: 2.40) without negation.

In Experiment 4, for the filler group with the number manipulation, a mean rating of 5.35 (standard deviation: 1.98) was found in the singular condition and 4.45 (standard deviation: 2.02) in the plural condition. For minimizers, a mean rating of 5.84 (standard deviation: 1.82) was found with negation and 3.51 (standard deviation: 2.18) without negation.

The results show that participants are sensitive to deviations from the standard form of idioms and it is reflected clearly in their ratings whether grammatical requirements such as licensing by negation are satisfied or not. This suggests that participants did not indifferently accept any kind of transformation when it comes to rating sentences containing idiomatic expressions. We conclude from this that the ratings for the syntactic manipulations give us a reliable impression of how acceptable they are for speakers.

In Experiment 3, two additional filler items with sentences in passive voice were included for exploratory purposes. A detail that can be noticed about the main results of Experiment 3 with respect to the passive construction is that the two idioms for which it is judged as absolutely unacceptable both mean 'to die' (das Zeitliche segnen, den Löffel abgeben); they are thus the only ones whose literal paraphrase would be an unaccusative verb (which cannot be easily passivized). The two filler items in Experiment 3 were included with the purpose of potentially checking whether this particular property (unaccusative paraphrase) might have influenced the results. The filler items included passivized transitive verbs that also express 'to die', but are less idiomatic, namely *den Tod finden* 'to find death' and *das Leben lassen* 'to lose life'. For these items, we indeed also found rating close to floor level for the passivized version (mean rating: 1.68). This suggests that for further research on the passivizability of idioms, the passivizability of the literal paraphrase might play a role and should be systematically controlled.

APPENDIX B: ADDITIONAL INFORMATION ON THE STATISTICAL ANALYSIS

B.1 *The motivation for focusing on interactions*

In the analysis of our experiments, we focus on the question whether the factors COMPOSITIONALITY and CONTEXT show a significant interaction with the factor STRUCTURE. The reasoning behind this is the following: let us assume that we find a significant difference between compositional and non-compositional idioms in one of the marked structure, e.g., in left dislocation, as illustrated in Figure 1.

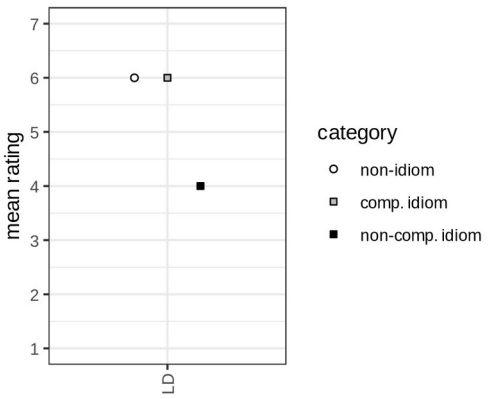

Figure 1: Hypothetical results for left dislocation

In this situation, the crucial question is whether the observed difference between compositional and non-compositional idioms is really due to the fact that left-dislocation of non-compositional idioms is less acceptable than left dislocation of compositional ones. An alternative possibility is that sentences containing non-compositional idioms are generally judged as less acceptable for unrelated reasons (e.g., due to lower familiarity, frequency, etc.). As we deliberately picked our idioms sample from a limited set (idioms consisting of a transitive verb and a definite DP) in order to make sure that they can occur in all syntactic structures that we intended to test, it was difficult to fully eliminate these potential confounds by controlling/matching the items for all non-syntactic properties. Thus, in order to find out whether it is less acceptable to left-dislocate non-compositional idioms than compositional ones, it is not sufficient to look at the difference between these two groups. It is crucial to consider a *difference in differences*, viz., compare the difference between non-compositional and compositional in sentences with left dislocation to the difference between non-compositional and compositional in sentences with canonical word order. If the difference between non-compositional and composi-

tional is significantly larger in a marked structure like left dislocation than it is in a canonical sentence as illustrated in Figure 2, we can conclude that compositionality influences the degree to which an idiom can be left-dislocated; if the difference between compositional and non-compositional is similar in LD and canonical word order as in Figure 3, we cannot draw such a conclusion.

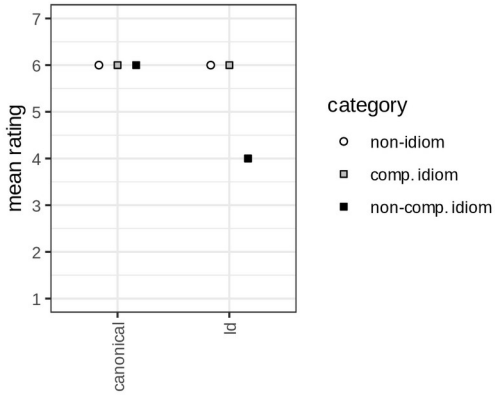

Figure 2: Hypothetical results for left dislocation in comparison to baseline

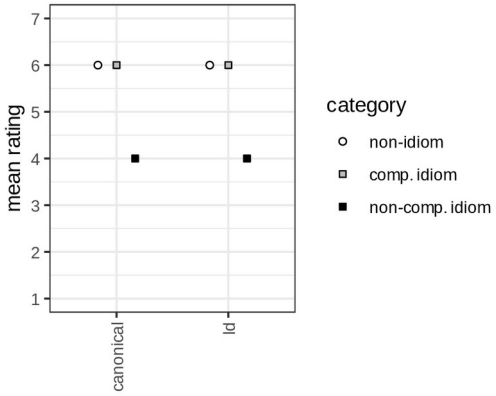

Figure 3: Hypothetical results for left dislocation in comparison to baseline

In our statistical analysis, a *difference in differences* corresponds to a significant interaction between COMPOSITIONALITY and STRUCTURE. Analogously, this reasoning applies to the factor CONTEXT, for which it is also crucial to look at its interaction with STRUCTURE.

Note that the presumed identical behavior of compositional idioms and non-idioms in Figures 1-3 is based on the idealized assumption that we are able to pick idioms that are 100% compositional. Since this will not necessarily be the case, we include non-idiomatic items for comparison. We expect that higher restrictiveness of a structure will mainly be reflected in a difference between compositional and non-compositional id-

ioms, but there could also be an informative difference between non-idioms and compositional idioms.

B.2 *Model outputs*

The following tables show the model results for the fixed effects in Experiments 1-4.

For the factor CONSTRUCTION, treatment contrast coding was used, with canonical as the baseline. In the output, a term like "can-pre" thus represents the comparison between canonical and prefield. Abbreviations: can = canonical; pre = prefield; ld = left dislocation; scr = scrambling; ana = anaphor; nom = nominalization (without "of"); nof = nominalization with "of"; pas = passive, cle = cleftlike.

For the factor CONTEXT, sum coding was used. In the output, the term context1 represents the comparison between broad focus and the overall mean, and context2 represents the comparison between polarity focus and the overall mean.

For the factor COMPOSITIONALITY, forward difference coding was used. In the output, the term comp2-1 represents the difference between non-idioms and compositional idioms, and the term comp3-2 represents the difference between compositional idioms and non-compositional idioms.

The following significance codes are used in the tables: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, . for $p < 0.1$.

The parsimonious model that we identified for Experiment 1 included random intercepts for participant and item, as well as the following by-participant random slopes: construction.can-ana:comp2-1, construction.can-ld:comp2-1, comp3-2, comp2-1, construction.can-ana, construction.can-scr, construction.can-ld, construction.can-pre, and

the following by-item random slopes: construction.can-scr, construction.can-pre. The output is shown in Table 1.

Table 1: Model output (fixed effects) for Experiment 1

	Estimate	SE	t-value	p-value	
(Intercept)	6.17	0.117	52.863	< 0.001	***
construction.can-pre	-1.274	0.119	-10.715	< 0.001	***
construction.can-ld	-1.917	0.189	-10.128	< 0.001	***
construction.can-scr	-0.986	0.162	-6.079	< 0.001	***
construction.can-ana	-1.235	0.131	-9.432	< 0.001	***
context1	0.16	0.1	1.599	0.116	
comp2-1	-0.266	0.192	-1.386	0.174	
comp3-2	-0.001	0.189	-0.007	0.994	
construction.can-pre:context1	-0.931	0.111	-8.403	< 0.001	***
construction.can-ld:context1	-0.715	0.189	-3.777	< 0.001	***
construction.can-scr:context1	-0.678	0.112	-6.069	< 0.001	***
construction.can-ana:context1	-0.531	0.131	-4.053	< 0.001	***
construction.can-pre:comp2-1	0.251	0.188	1.336	0.191	
construction.can-ld:comp2-1	0.092	0.158	0.58	0.563	
construction.can-scr:comp2-1	-0.074	0.327	-0.226	0.824	
construction.can-ana:comp2-1	0.048	0.169	0.287	0.775	
construction.can-pre:comp3-2	-0.957	0.188	-5.096	< 0.001	***
construction.can-ld:comp3-2	-0.562	0.155	-3.629	< 0.001	***
construction.can-scr:comp3-2	-0.897	0.327	-2.743	0.012	*
construction.can-ana:comp3-2	-0.601	0.155	-3.879	< 0.001	***
context1:comp2-1	-0.085	0.124	-0.682	0.496	
context1:comp3-2	0.036	0.119	0.308	0.759	
construction.can-pre:context1:comp2-1	0.039	0.155	0.25	0.803	
construction.can-ld:context1:comp2-1	-0.249	0.158	-1.58	0.116	
construction.can-scr:context1:comp2-1	0.047	0.155	0.306	0.759	
construction.can-ana:context1:comp2-1	-0.005	0.169	-0.027	0.978	
construction.can-pre:context1:comp3-2	-0.184	0.155	-1.187	0.235	
construction.can-ld:context1:comp3-2	-0.07	0.155	-0.449	0.653	
construction.can-scr:context1:comp3-2	-0.147	0.155	-0.949	0.343	
construction.can-ana:context1:comp3-2	-0.154	0.155	-0.993	0.321	

The parsimonious model that we identified for Experiment 2 included random intercepts for participant and item, as well as the following by-participant random slopes: construction.can-scr:comp3-2, construction.can-ana:comp2-1, construction.can-pre:comp2-1, comp3-2, comp2-1, construction.can-ana, construction.can-scr, construction.can-ld,

construction.can-pre, and the following by-item random slopes: construction.can-scr:context1, construction.can-pre:context1, context1, construction.can-scr, construction.can-pre. The output is shown in Table 2.

Table 2: Model output (fixed effects) for Experiment 2

	Estimate	SE	t-value	p-value	
(Intercept)	5.719	0.141	40.696	< 0.001	***
construction.can-pre	-1.311	0.144	-9.13	< 0.001	***
construction.can-ld	-1.892	0.159	-11.902	< 0.001	***
construction.can-scr	-0.947	0.165	-5.755	< 0.001	***
construction.can-ana	-0.917	0.192	-4.782	< 0.001	***
context1	0.511	0.136	3.765	< 0.001	***
comp2-1	-0.013	0.158	-0.079	0.937	
comp3-2	0.008	0.156	0.054	0.958	
construction.can-pre:context1	-1.114	0.139	-8.042	< 0.001	***
construction.can-ld:context1	-1.028	0.159	-6.467	< 0.001	***
construction.can-scr:context1	-0.742	0.128	-5.789	< 0.001	***
construction.can-ana:context1	-0.633	0.192	-3.304	0.002	**
construction.can-pre:comp2-1	-0.317	0.198	-1.597	0.12	
construction.can-ld:comp2-1	-0.183	0.157	-1.17	0.242	
construction.can-scr:comp2-1	-0.133	0.319	-0.418	0.68	
construction.can-ana:comp2-1	-0.217	0.182	-1.189	0.237	
construction.can-pre:comp3-2	-0.438	0.193	-2.267	0.031	*
construction.can-ld:comp3-2	-0.571	0.157	-3.644	< 0.001	***
construction.can-scr:comp3-2	-0.975	0.323	-3.02	0.006	**
construction.can-ana:comp3-2	-0.654	0.157	-4.176	< 0.001	***
context1:comp2-1	-0.162	0.13	-1.25	0.214	
context1:comp3-2	-0.092	0.128	-0.718	0.474	
construction.can-pre:context1:comp2-1	0.008	0.175	0.048	0.962	
construction.can-ld:context1:comp2-1	-0.025	0.157	-0.16	0.873	
construction.can-scr:context1:comp2-1	0.067	0.195	0.343	0.734	
construction.can-ana:context1:comp2-1	0.117	0.182	0.64	0.523	
construction.can-pre:context1:comp3-2	0.079	0.169	0.468	0.642	
construction.can-ld:context1:comp3-2	-0.046	0.157	-0.293	0.77	
construction.can-scr:context1:comp3-2	0.042	0.2	0.208	0.837	
construction.can-ana:context1:comp3-2	0.004	0.157	0.027	0.979	

The parsimonious model that we identified for Experiment 3 included random intercepts for participant and item, as well as the following by-participant random slopes: construction.can-nom:comp2-1, construction.can-pas:comp2-1, comp2-1, construction.can-

wh, construction.can-nom, construction.can-pas, construction.can-ld, construction.can-pre, and the following by-item random slopes: construction.can-wh, construction.can-nom, construction.can-pas, construction.can-pre. The output is shown in Table 3.

Table 3: Model output (fixed effects) for Experiment 3

	Estimate	SE	t-value	p-value	
(Intercept)	5.742	0.162	35.410	< 0.001	***
construction.can-pre	-0.356	0.125	-2.856	0.009	**
construction.can-ld	-1.547	0.219	-7.068	< 0.001	***
construction.can-pas	-1.75	0.227	-7.723	< 0.001	***
construction.can-nom	-2.606	0.238	-10.940	< 0.001	***
construction.can-wh	-2.503	0.286	-8.757	< 0.001	***
comp2-1	-0.217	0.296	-0.730	0.470	
comp3-2	-0.017	0.288	-0.058	0.954	
construction.can-pre:comp2-1	-0.2	0.271	-0.739	0.465	
construction.can-ld:comp2-1	-0.016	0.245	-0.067	0.947	
construction.can-pas:comp2-1	-0.1	0.497	-0.200	0.843	
construction.can-nom:comp2-1	-0.7	0.446	-1.569	0.13	
construction.can-wh:comp2-1	-1.591	0.44	-3.615	0.002	**
construction.can-pre:comp3-2	-0.717	0.281	-2.653	0.012	*
construction.can-ld:comp3-2	-0.883	0.244	-3.601	< 0.001	***
construction.can-pas:comp3-2	-1.1	0.496	-2.245	0.035	*
construction.can-nom:comp3-2	-0.467	0.44	-1.06	0.301	
construction.can-wh:comp3-2	-0.575	0.44	-1.307	0.205	

The parsimonious model that we identified for Experiment 4 included random intercepts for participant and item, as well as the following by-participant random slopes: construction.can-nom:comp3-2, construction.can-ana:comp3-2, construction.can-cle:comp3-2, construction.can-nom:comp2-1, construction.can-ana:comp2-1, construction.can-cle:comp2-1, comp3-2, comp2-1, construction.can-nom, construction.can-ana, construction.can-cle, construction.can-pas, construction.can-nof , and the following by-item random slope: construction.can-pas. The output is shown in Table 4.

Table 4: Model output (fixed effects) for Experiment4

	Estimate	SE	t-value	p-value	
(Intercept)	5.086	0.237	21.422	< 0.001	***
construction.can-nof	-0.619	0.199	-3.107	0.005	**
construction.can-pas	-0.894	0.166	-5.399	< 0.001	***
construction.can-cle	-0.836	0.149	-5.606	< 0.001	***
construction.can-ana	-0.156	0.133	-1.17	0.252	
construction.can-nom	-0.928	0.219	-4.233	< 0.001	***
comp2-1	-0.617	0.287	-2.151	0.036	*
comp3-2	-0.608	0.288	-2.115	0.039	*
construction.can-nof:comp2-1	0.217	0.256	0.847	0.397	
construction.can-pas:comp2-1	-0.1	0.329	-0.304	0.763	
construction.can-cle:comp2-1	-1	0.257	-3.888	< 0.001	***
construction.can-ana:comp2-1	0	0.267	0	1	
construction.can-nom:comp2-1	0.533	0.283	1.882	0.065	.
construction.can-nof:comp3-2	0.133	0.256	0.521	0.602	
construction.can-pas:comp3-2	-0.533	0.329	-1.622	0.115	
construction.can-cle:comp3-2	0.117	0.262	0.445	0.658	
construction.can-ana:comp3-2	-0.217	0.265	-0.818	0.416	
construction.can-nom:comp3-2	0.45	0.291	1.547	0.128	

B.3 *Post-hoc comparisons*

We ran a post-hoc analysis for each experiment to test if the syntactic structures differed from each other with respect to the interaction with the factor COMPOSITIONALITY.

For Experiments 1 and 2, there were twelve additional comparisons. We thus set the alpha-level to 0.05/12 (0.0042) to compensate for the higher likelihood of erroneous inferences in multiple testing (Bonferroni correction). The results are shown in Tables 5 and 6.

For Experiments 3 and 4, there were 20 comparisons and the Bonferroni-adjusted alpha-level was thus set to 0.05/20 (0.0025). The results are shown in Tables 7 and 8.

Table 5: post-hoc comparison of the STRUCTURE × COMPOSITIONALITY *interaction at individual levels of the factor* STRUCTURE *in Experiment 1*

	non-idiom / comp. idiom	comp. idiom / non-comp. idiom
prefield / LD	t = -0.84, p = 0.41 (n.s.)	t = 2.10, p = 0.04 (n.s.)
prefield / scrambling	t = -0.94, p = 0.36 (n.s)	t = 0.17, p = 0.86 (n.s.)
prefield / anaphor	t = -1.01, p = 0.32 (n.s.)	t = 1.90, p = 0.07 (n.s.)
LD / scrambling	t = -0.50, p = 0.62 (n.s.)	t = -1.02, p = 0.32 (n.s.)
LD / anaphor	t = -0.25, p = 0.80 (n.s.)	t = -0.25, p = 0.80 (n.s.)
scrambling / anaphor	t = 0.37, p= 0.72 (n.s.)	t = 0.91, p = 0.38 (n.s.)

Table 6: post-hoc comparison of the STRUCTURE × COMPOSITIONALITY *interaction at individual levels of the factor* STRUCTURE *in Experiment 2*

	non-idiom / comp. idiom	comp. idiom / non-comp. idiom
prefield / LD	t = 0.67, p = 0.51 (n.s.)	t = -0.69, p = 0.49 (n.s.)
prefield / scrambling	t = 0.54, p = 0.60 (n.s)	t = -1.57, p = 0.13 (n.s.)
prefield / anaphor	t = 0.46, p = 0.65 (n.s.)	t = -1.12, p = 0.27 (n.s.)
LD / scrambling	t = 0.16, p = 0.88 (n.s.)	t = -1.25, p = 0.22 (n.s.)
LD / anaphor	t = -0.18, p = 0.86 (n.s.)	t = -0.53, p = 0.59 (n.s.)
scrambling / anaphor	t = -0.25, p = 0.80 (n.s.)	t = 0.99, p = 0.33 (n.s.)

Table 7: post-hoc comparison of the STRUCTURE × COMPOSITIONALITY *interaction at individual levels of the factor* STRUCTURE *in Experiment 3*

	non-idiom / comp. idiom	comp. / non-comp. idiom
prefield / LD	t = 0.68, p = 0.50 (n.s.)	t = -0.62, p = 0.54 (n.s.)
prefield / passive	t = 0.20 p = 0.85 (n.s)	t = -0.76, p = 0.45 (n.s.)
prefield / nominalization	t = -1.09, p = 0.29 (n.s.)	t = 0.55, p = 0.59 (n.s.)
prefield / which-question	t = -3.06, p = 0.006 (n.s.)	t = 0.31, p = 0.76 (n.s.)
LD / passive	t = -0.17, p = 0.87 (n.s.)	t = -0.44, p = 0.66 (n.s.)

LD / nominalization	t = -1.53, p = 0.14 (n.s.)	t = 0.95, p = 0.35 (n.s.)
LD / which-question	t = -3.58, p = 0.0017 (sign.)	t = 0.70, p = 0.49 (n.s.)
passive / nominalization	t = -0.97, p = 0.34 (n.s.)	t = 2.25, p = 0.02 (n.s.)
passive / which-question	t = -2.42, p = 0.02 (n.s.)	t = 1.04, p = 0.31 (n.s.)
nom. / which-question	t = -1.55, p = 0.13 (n.s.)	t = -0.19, p = 0.85 (n.s.)

Table 8: post-hoc comparison of the STRUCTURE × COMPOSITIONALITY *interaction at individual levels of the factor* STRUCTURE *in Experiment 4*

	non-idiom / comp. idiom	comp. / non-comp. idiom
nom. with "of" / passive	t = -0.96, p = 0.34 (n.s.)	t = -2.03, p = 0.05 (n.s.)
nom. with "of" / cleftlike	t = -4.73, p < 0.001 (sign.)	t = -0.06, p = 0.95 (n.s.)
nom. with "of" / anaphor	t = -0.81, p = 0.42 (n.s.)	t = -1.32, p = 0.19 (n.s.)
nom. with "of" / nom. without "of"	t = 1.12, p = 0.27 (n.s.)	t = 1.09, p = 0.28 (n.s.)
passive / cleftlike	t = -2.73, p = 0.01 (n.s.)	t = 1.95, p = 0.06 (n.s.)
passive / anaphor	t = 0.77, p = 0.77 (n.s.)	t = 0.94, p = 0.35 (n.s.)
passive / nom. without "of"	t = 1.81, p = 0.08 (n.s.)	t = 2.76, p = 0.0096 (n.s.)
cleftlike / anaphor	t = 3.73, p < 0.001 (sign.)	t = -1.23, p = 0.23 (n.s.)
cleftlike / nom. without "of"	t = 5.39, p < 0.001 (sign.)	t = 1.12, p = 0.27 (n.s.)
anaphor / nom. without "of"	t = 1.82, p = 0.08 (n.s.)	t = 2.23, p = 0.03 (n.s.)

APPENDIX C: ADDITIONAL EXPLORATORY ANALYSES

C.1 *By-participant analysis*

We present data from our first experiment (polarity focus context) here to illustrate the by-participant distribution in Figure 4. By-participant analyses of the other experiments are available in the OSF repository.

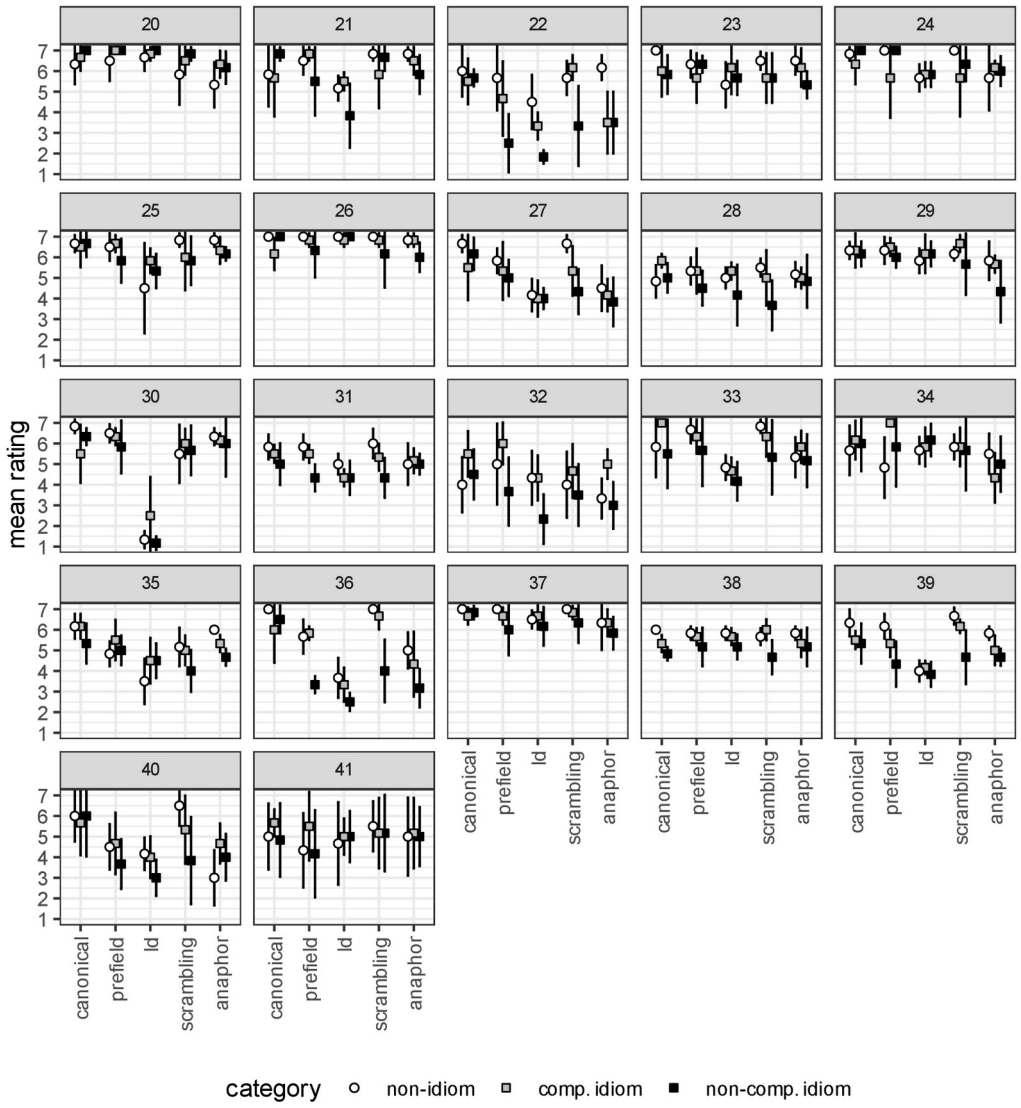

Figure 4: By-participant results for the group that saw the polarity context in Exp. 1.

Based on visual inspection, there is considerable inter-speaker variation with respect to the effect size of the factor compositionality. Focusing on those participants who show the largest acceptability differences in this respect (#22, #32, #36, #39), we can observe that even for these speakers, there do not seem to be structures that are not possible with idioms at all. The observation that compositional idioms are similarly acceptable as non-idioms also holds for these speakers, and the larger decrease only concerns the non-

compositional idioms. In particular, for each participant there is a subset of idioms that can undergo scrambling (namely at least the set that we categorized as compositional). We interpret this as evidence that scrambling of idiom parts is not completely excluded for speakers of German.

C.2 *Relation between selected fillers and target items*

As discussed in Appendix A, some of our non-critical stimuli were included with the intention to check for a potential influence of the participants' willingness to engage in plays on words / language games. It is conceivable that this is a factor that could lead participants to treat idioms as more similar to non-idioms (e.g., by considering the literal interpretation for their judgment, even though it would be odd/funny in the provided context), thus reducing the difference between the categories and potentially masking effects in our experiments. In our view, the stimuli including a pun-like coordination of idiomatic and non-idiomatic meaning in Experiment 1 and the stimuli with a number manipulation in Experiments 2-4 are particularly likely to reflect this individual tendency. In an additional exploratory analysis, we therefore included each participant's mean rating for these groups of stimuli as a predictor in a linear mixed model. In all experiments, we consistently found the following: the higher a participant rated these stimuli, the higher their rating was for some of the marked structures in comparison to the baseline (significant increase for prefield, LD, scrambling, anaphor in Exp. 1; anaphor in Exp. 2; passive, nominalization, which-question in Exp. 3; nominalization with "of" in Exp. 4). But in neither of the experiments did this factor interact significantly with compositionality, i.e., it did not specifically affect the difference between non-idioms and idioms, or between compositional/non-compositional idioms. These

analyses thus do not provide support for the idiom-specific assumption above. The detailed results of these analyses can be found in the OSF repository.