GLOSSA psycholinguistics

Bader, M., & Meng, M. (2023). Processing noncanonical sentences: Effects of context on online processing and (mis)interpretation. *Glossa Psycholinguistics* 2(1): 1, pp. 1–45. DOI: https://doi.org/10.5070/G6011117

eScholarship University of California

Processing noncanonical sentences: Effects of context on online processing and (mis)interpretation

Markus Bader, Goethe University Frankfurt, DE, bader@em.uni-frankfurt.de

Michael Meng, Merseburg University of Applied Sciences, DE, michael.meng@hs-merseburg.de

Prior research has shown that sentences with noncanonical argument order (e.g., patient-before-agent instead of agent-before-patient order) are associated with additional online processing difficulty, but that this difficulty can be alleviated if the discourse context licenses noncanonical order. Other studies demonstrated that noncanonical sentences are prone to misinterpretation effects: comprehenders sometimes seem to form interpretations with incorrect assignments of semantic roles to argument NPs. However, those studies tested noncanonical sentences in isolation. To further clarify the source of misinterpretation effects, we designed three experiments that investigated how discourse properties licensing noncanonical order affect online processing and final interpretation. All experiments tested unambiguous active declarative sentences in German with agentive verbs and two arguments, probing both online processing difficulty (using selfpaced reading) and accuracy of interpretation (using wh-comprehension questions). Besides word order (subject-before-object, SO vs. object-before-subject, OS), we varied the context preceding the target sentence (neutral context vs. context licensing OS, Experiment 1), the type of NP serving as object (definite vs. demonstrative NP, Experiment 2) and the type of question probing comprehension (two-argument vs. one-argument wh-questions, Experiment 3). Consistent with earlier findings, we observed that discourse properties licensing OS order facilitated online processing in early sentence regions. However, they did barely affect accuracy on comprehension questions, with accuracy instead being a function of word order and question type. Our results support models that explain misinterpretation effects in terms of task-specific retrieval processes. A retrieval mechanism capturing the effects of question type is proposed.

1. Introduction

Sentences with noncanonical argument order have provided a long-standing puzzle for theories of human sentence comprehension. Passive clauses like *The athlete was called by the trainer* or object clefts like *It was the athlete that the trainer called* deviate from the canonical agent-before-patient order by assigning the patient role to the first NP and the agent role to the second NP. Two findings concerning the processing of sentences with noncanonical argument order have provided a particular challenge. First, noncanonical sentences are often associated with additional processing cost (e.g., Bader & Meng, 1999; Hopp et al., 2020; Kaiser & Trueswell, 2004; Koizumi & Imamura, 2017; Schlesewsky et al., 2000; Yano & Koizumi, 2018). This is reflected in online and offline measures such as increased reading times, specific ERP patterns, or lower acceptability ratings. Starting with Ferreira (2003), particular attention has been devoted to the final interpretation of noncanonical sentences. Studies assessing offline comprehension demonstrated that noncanonical sentences are prone to misinterpretation effects. When comprehenders process noncanonical sentences, they sometimes seem to arrive at interpretations with incorrect assignments of semantic roles to argument NPs (Bader & Meng, 2018; Ferreira, 2003; Gibson et al., 2013; Paolazzi et al., 2019, among others).

Two different types of accounts have been proposed for misinterpretation effects with noncanonical sentences. According to the good-enough model of sentence comprehension (henceforth referred to as GE model, Christianson, 2016; Ferreira & Patson, 2007; Karimi & Ferreira, 2016), misinterpretation errors reflect the application of heuristic strategies that are assumed to operate in parallel to algorithmic processing routines, including heuristics favoring the agent-before-patient over patient-before-agent order, and plausible over implausible interpretations (see Townsend & Bever, 2001 for a related proposal). Alternative accounts have ascribed misinterpretation effects to processes involved in maintaining or operating on memory representations for sentences (Bader & Meng, 2018; Cutter et al., 2022; Meng & Bader, 2021; Paolazzi et al., 2019). For example, Bader and Meng (2018) proposed that misinterpretations reflect processes of memory retrieval required by the agent/patient naming task which was used to assess comprehension in Ferreira (2003) and in Experiment 1 of Bader and Meng (2018).

The current study attempts to shed further light on the processing of noncanonical sentences by examining online processing and offline comprehension in parallel. The relation between online and offline measures has received considerable attention in research on another type of misinterpretation effect: lingering misinterpretations in garden path sentences such as *While Anna dressed the baby that was cute and cuddly played in the crib*. The central finding obtained by studies combining online and offline measures is that misinterpretation effects – as measured by offline comprehension tasks – persist even when online measures provide clear signs of syntactic reanalysis, such as increased reading times in the disambiguating region of garden path sentences (Christianson et al., 2010; Dempsey & Brehm, 2020; Qian et al., 2018). Furthermore,

comprehension accuracy was found to be independent from reading times in the disambiguating regions (Christianson & Luke, 2011; Chromý, 2021; Wonnacott et al., 2016).

In contrast, only a few studies on comprehending noncanonical sentences have related measures of online processing difficulty to measures of final interpretation, and the studies we are aware of are restricted to passives (Grillo et al., 2018; Paolazzi et al., 2019, 2021). A main finding of these studies is that passive sentences lead to (moderate) misinterpretation effects. Reading times for trials with correct answers to comprehension questions, however, indicated that passives were consistently read faster than active controls, with the reading time advantage showing up on the past participle and the following by-phrase. As the authors argue, the reading time advantage for correctly interpreted passives is unexpected under a GE account, as a correct interpretation of noncanonical sentences presupposes an algorithmic analysis, which should require more effort than a possibly wrong analysis relying on fast but error prone heuristics. Instead, an analysis in terms of surprisal theory (Levy, 2008) is proposed, based on the observation that passives include cues such as the passive auxiliary, the past participle and the preposition *by* which enable more specific predictions towards upcoming material. The slightly higher error rates with passives on comprehension questions are assumed to be unrelated to online processing mechanisms, but rather due to greater difficulty in maintaining the representation for passives in working memory and using this representation in the process of answering the comprehension question.

Our study extends this line of research by looking at a different type of noncanonical sentences with patient-before-agent order: active declarative sentences with the object in sentence-initial position, henceforth referred to as OS sentences. Whereas passives only reverse the canonical order of thematic roles, OS sentences also reverse the order of syntactic functions, thereby deviating from the more common subject-before-object order. This is shown in (1) and (2) for German, the language under investigation in this paper.

(1) Der Athlet ist vom Trainer angerufen worden. *passive*
the.NOMINATIVE athlete has by-the.DATIVE trainer called been
'The athlete was called by the trainer.'

(2) Den Athlet hat der Trainer angerufen. *active OS*
the.ACCUSATIVE athlete has the.NOMINATIVE trainer called
'The trainer called the athlete.'

German is a language with relatively free word order. Subject and object can appear in either SO or OS order. For simple declarative clauses with agentive verbs as used in our experiments, SO order is unmarked, whereas the use of OS order is marked and subject to discourse constraints (see Frey, 2004; Lenerz, 1977, among others). For example, fronting the object to sentence-initial position as in (2) is licensed when the object referent is focused due to a

preceding wh-question (Fanselow et al., 2008) or when it is given, either directly by having been mentioned in the preceding context (Bader & Portele, 2021) or indirectly by standing in a partially ordered set relation to a referent of the preceding context (Weskott et al., 2011). In addition to the discourse status of the object (e.g., new or given, focused or not), the particular referential expression of the object (e.g., definite or demonstrative NP) also has a strong effect on the choice between SO and OS order, as will be explained in detail in the introduction to Experiment 2.

Prior research has demonstrated facilitating effects of discourse context on processing OS sentences in Finnish, German, Japanese and other languages. When presented in a licensing context, overall acceptability of OS sentences improved (Bornkessel & Schlesewsky, 2006; Burmester et al., 2014; Imamura et al., 2016) and online processing difficulty was reduced (Bornkessel et al., 2003; Kaiser & Trueswell, 2004; Koizumi & Imamura, 2017; Meng et al., 1999; Weskott et al., 2011). However, whether a licensing discourse context also affects (mis)interpretation of noncanonical sentences is less clear. While many studies included comprehension questions to make sure that participants read carefully, only a few attempted to systematically assess the content of the final interpretation, and the studies we are aware of tested sentences that were locally ambiguous between an SO and an OS reading. Kristensen et al. (2014) examined locally ambiguous SO and OS sentences in Danish. Target sentences followed a neutral context or a supportive context motivating (but not requiring) OS order. Selfpaced reading times showed a disadvantage for OS sentences irrespective of context type. However, yes/no comprehension questions testing thematic role assignment showed an improvement in comprehension accuracy for OS sentences in supportive contexts (rising from 51% to 75%), whereas accuracy for SO sentences was equally high for both contexts (about 91%). A similar effect of context on comprehension accuracy was reported in Vos and Friederici (2003) for locally ambiguous OS sentences in German.

Here, we report three experiments that investigated the processing of unambiguous German active SO and OS sentences in context, probing both online difficulty (using selfpaced reading) and sentence-final interpretation (using wh-questions). Experiment 1 manipulated properties of the discourse context. Target sentences were presented either in a neutral context or in a context licensing OS order by establishing the referent of the object NP as given, as shown in (3).

(3) a. *Context sentence*

Neutral Heute hat es viel Unruhe vor und während des Unterrichts gegeben.
'Today there was a lot of unrest before and during class.'

Supportive Der neue Lehrer war mal wieder nicht zufrieden mit einem Schüler aus der fünften Klasse.
Once again, the new teacher was not satisfied with a student from the 5th grade class.

b. *Target sentence*

SO Der Lehrer hat den Schüler deshalb vor der Klasse angeschrien.
the teacher has the student therefore in-front-of the class shouted-at
'The teacher therefore yelled at the student in front of the class.'

OS Den Schüler hat der Lehrer deshalb vor der Klasse angeschrien.
the student has the teacher therefore in-front-of the class shouted-at
'The teacher therefore yelled at the student in front of the class.'

Experiment 2 used only contexts licensing OS order but varied the referential form of the object of the target sentence – the object was either a demonstrative or a definite NP. Experiment 3 is a follow up to Experiments 1 and 2 using less complex comprehension questions to re-assess sentence interpretation.

Our experiments address two main questions. The first question relates to the effect of context and referential form on the final interpretation comprehenders arrive at. As discussed above, prior research has demonstrated that a licensing context can reduce online processing difficulty for OS sentences. Our experiments aim to find out whether a context and referential form supporting OS order also affects misinterpretation errors. Studies capitalizing on comprehenders' final interpretation have presented OS sentences out of context, which may have contributed to the high error rates observed for English OS cleft-sentences (Ferreira, 2003) and German OS main clauses (Bader & Meng, 2018; Meng & Bader, 2021). For garden-path sentences, context has been argued to strengthen misinterpretation effects if the initial analysis integrates well with the preceding context (Christianson & Luke, 2011; Dempsey & Brehm, 2020). If this also holds for the processing of unambiguous noncanonical OS sentences, misinterpretation effects should persist or become even stronger, as our manipulations of context and referential form license, but do not force, OS order and are also compatible with SO order. Alternatively, it could be argued – following Karimi and Ferreira (2016) – that the human parser puts more effort on algorithmic processing if context and referential form are not only compatible with SO order, but with OS order as well. In such a situation, it is less certain that the initial analysis arising from the agent-before-patient heuristic is correct, and a more detailed analysis with fewer misinterpretation errors is attempted.

Our second question concerns the underlying source of misinterpretation errors that have been found for OS sentences. Do such errors result from the application of heuristics during online parsing or are they caused by faulty memory retrievals triggered by a cue or a comprehension question following the sentence? Our experiments address this question in two ways: by analyzing the relationship between reading times and comprehension accuracy and by using a more natural task to assess comprehension.

Regarding the relationship between reading times and comprehension accuracy, we will specifically examine how reading times for noncanonical OS sentences depend on whether comprehension questions are answered correctly or not. A GE account leads us to expect that

reading times for OS sentences should be faster when they are misinterpreted. According to the model proposed in Karimi and Ferreira (2016), heuristic and algorithmic processing routines are launched simultaneously. Being based on simple rules or sentence templates (Townsend & Bever, 2001), the heuristic routines can generate an interim result faster. However, in case of noncanonical sentences, applying rules like the agent-before-patient heuristic leads to wrong interpretations. To arrive at the correct interpretation, the output of the heuristic route has to be reconciled with a representation based on algorithmic processing routines, which are assumed to operate more slowly, but typically deliver the correct interpretation. The increased effort required to complete algorithmic processing should therefore be associated with higher reading times for trials with the comprehension question answered correctly. Under an account that does not ascribe misinterpretations to parsing errors, a systematic difference in reading times for OS sentences depending on whether comprehension questions were answered correctly or not is not expected.

As a further way to uncover the underlying source of misinterpretation errors, our study extends previous work in using a more natural task to assess comprehension. Ferreira (2003) used the agent-patient naming task for this purpose. After hearing a sentence, participants are presented a cue word like DO-er or UNDERGOER and have to name the corresponding argument from the preceding sentence. Agent-patient naming was also used in Experiment 1 of Bader and Meng (2018). The agent-patient naming task is unnatural insofar as participants in real-world settings are hardly ever required to respond to cue words like DO-er or UNDERGOER. The plausibility judgment task used in Experiment 2 of Bader and Meng (2018) is unnatural as well as it is a meta-linguistic task.

To probe comprehension, the current study uses wh-questions asking for the agent or patient of the preceding sentence. Wh-questions in Experiments 1 and 2 contained two arguments, agent and patient. Experiment 3 examines the influence of the complexity of the comprehension question by presenting questions that are simpler and contain one argument only, agent or patient.

2. Experiment 1

Experiment 1 examined the processing of simple declarative active sentences in German with either subject or object in sentence-initial position (SO vs. OS order). Sentences contained agentive verbs and were presented in a neutral context or in a context that was supportive in the sense that it licensed OS order. The experiment used selfpaced reading with word-by-word, non-cumulative presentation to assess online processing difficulty. Each target sentence was followed by a wh-question to assess readers' final interpretation.

Experiment 1 builds on the insight that fronting an object to sentence-initial position is licensed when the referent of the object has been introduced before. This is the case in the supportive context condition, but not in the neutral context condition. SO sentences do not show

this kind of restriction. Hence, a supportive context makes a continuation using an OS sentence more probable. However, the context does not force an OS continuation and is also compatible with an SO sentence.

2.1 Method

2.1.1 Participants

Sixty students, 35 from Goethe University Frankfurt and 25 from Merseburg University of Applied Sciences, participated in Experiment 1. All were native speakers of German and naive with respect to the purpose of the experiment. Their mean age was 25.3, ranging from 19 to 54. Participants received €5 or course credit for participation.

2.1.2 Materials

We constructed 24 active declarative sentences containing an agentive verb describing a simple event involving an agent and a patient (see **Table 1** for an example). Each sentence occurred in four versions, according to the factors Order (SO vs. OS) and Context (neutral vs. supportive). Each target sentence was preceded by a context sentence. A neutral context sentence set up a scene without mentioning the subject or the object referent of the following target sentence. Neutral context sentences are assumed to be fully compatible with a following SO sentence but less so with a following OS sentence. A supportive context sentence already introduced both subject and object referent.

Table 1: A complete stimulus item for Experiment 1 including a patient-wh question.

Context:	Neutral	Wie jedes Jahr hat auch dieses Jahr wieder ein rauschendes Fest im Schloss stattgefunden.
		‘Like every year, a wonderful celebration took place at the castle this year.’
	Supportive	Der betagte König hat auf der großen Feier einen Botschafter in einem ausgefallenen Kostüm begrüßt.
		‘The old king greeted an ambassador in a fancy masquerade costume during the great celebration.”
Target:	SO	Der König hat den Botschafter dabei trotz der Maske erkannt.
		‘The king recognized the ambassador despite the mask.’
	OS	Den Botschafter hat der König dabei trotz der Maske erkannt.
		‘The king recognized the ambassador despite the mask.’
Question:		Wen hat jemand erkannt? – König – Botschafter
		‘Who did someone recognize? – King – Ambassador’

Each pair of context and target sentence was immediately followed by a wh-question to probe comprehension. Wh-questions either queried the agent or patient. To refer to the other argument of the target sentence, the indefinite pronoun *jemand/jemanden* ('someone.NOMINATIVE/.ACCUSATIVE') was used. Hence, a third factor was Question Type (agent-wh vs. patient-wh). This factor was varied between sentences. Half of the experimental sentences were followed by an agent question, the other half by a patient question.

The 24 experimental items were mixed with 48 fillers consisting of pairs of 2 sentences as well. All filler sentences were declarative sentences varying in syntactic structure, length and complexity.

2.1.3 Procedure

Participants were tested individually in our labs using the PsychoPy software (Peirce, 2007). Each experimental session consisted of two parts. In the first part, participants' reading span was assessed. The procedure and the results of the reading span task are not reported here.[1] After participants had completed the reading span task, the selfpaced reading task was started. Upon pressing the space bar, the two sentences of each trial appeared together on a single screen, with each letter replaced by an underline. When participants pressed the space bar again, the first word was shown at the position indicated by the underlines. With the next press of the space bar, the next word was shown and the first word was masked again. Using this procedure, participants were instructed to read through the sentences at their own pace.

When the last word was read, the comprehension question was displayed together with two answer options printed next to each other. For experimental sentences, the answer options referred to the agent and the patient of the target sentence. The correct answer was displayed equally often on the left and on the right side. Participants were instructed to select one answer by pressing the left or the right SHIFT key on the keyboard. The response deadline was set to 10 seconds. Participants did not get feedback concerning the correctness of their answer. After participants had selected an answer or the response deadline was reached, the next trial was initiated. Before starting the selfpaced reading task, participants received three practice trials.

2.2 Results

Statistical analyses were conducted using the R statistics software (R Core Team, 2020). Reading time and reaction time data were analyzed with linear mixed-effects models, accuracy data

[1] We originally planned to include a reading span test for all three experiments for reasons of comparison with our prior studies on misintepretation errors (Bader & Meng, 2018; Meng & Bader, 2021). Due to the COVID-19 pandemic, the reading span test could not be included in Experiments 2 and 3 and we therefore refrain from reporting any reading span results here.

were analyzed with generalized mixed-effects models, using the lme4 R-package (Bates, Mächler, et al., 2015). Models included the experimental factors and their interactions as fixed effects. Random effect terms were determined following the model fitting procedure proposed in Bates, Kliegl, et al. (2015). All factors were effect coded (0.5 vs. –0.5) and thus compared factor levels to each other. The questions for the filler sentences were answered with a mean accuracy of 98.9%. All participants had an accuracy of at least 83% and therefore no one was excluded from the analysis.

2.2.1 Reading times

The reading times were prepared for analysis in several steps, closely following Hofmeister (2011). First, extremely low or high reading times (rt $<$ 100 ms or rt $>$ 2500 ms) were removed. Second, reading times were log-transformed to avoid skewing. Third, residual reading times were computed from the log-transformed raw reading times, taking both experimental and filler items and contexts and target sentences into account. To obtain residual reading times, linear mixed-effects models were fitted to the log-transformed reading times with participants as random factor and the following fixed predictor variables: word length, the log-transformed position of the trial in the experiment, and the position of the word within the sentence. The latter factor was coded as a categorical variable with three levels: a word could either appear in sentence-initial position, in sentence-final position, or sentence-medial position (that is, in every linear position except the first and the final one). For analysis purposes, the target sentences were divided into several regions as shown in (4). In addition, the last word of the context sentence preceding the target sentence was analyzed (region S1-final). For each region, we report mean residual logarithmic reading times per word.

(4) Target sentence regions for reading time analysis

Der König	hat	den Botschafter	dabei trotz der Maske	erkannt.
The king	has	the ambassador	there despite the mask	recognized.
NP1	V-fin	NP2	midfield	V-end

Figure 1 shows the time course of reading the final word of the context sentence and the different regions of the target sentence. To make it easier to visualize the results, **Figure 2** shows the same mean reading times with a single plot for each region, including error bars. The results of the linear mixed-effects models for the individual regions are summarized in **Table 2**.

For the final word of the context sentence, a significant effect of Context was found due to higher reading times in supportive compared to neutral context sentences. On the initial NP of the target sentence (region NP1), reading times were again higher following supportive than following neutral contexts. Neither the factor Order nor the interaction between Order and Context were significant for NP1.

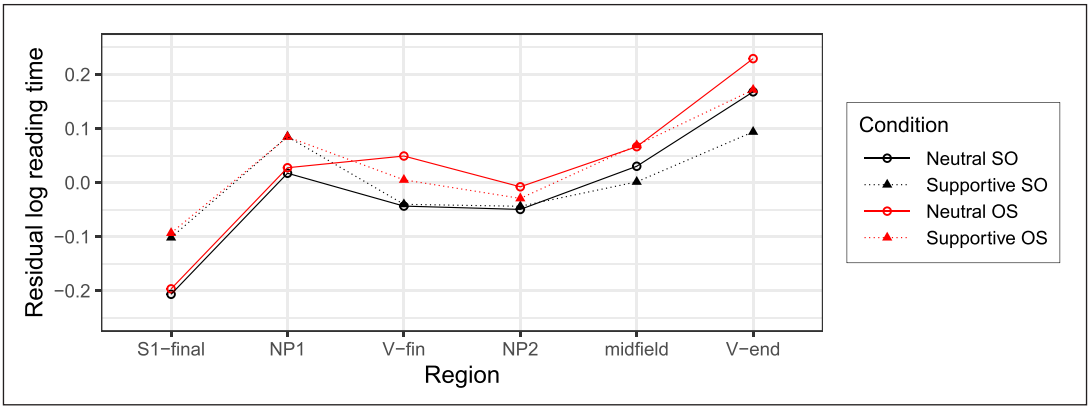


Figure 1: Mean reading times for the last word of the context sentence and the different regions of the target sentence in Experiment 1.

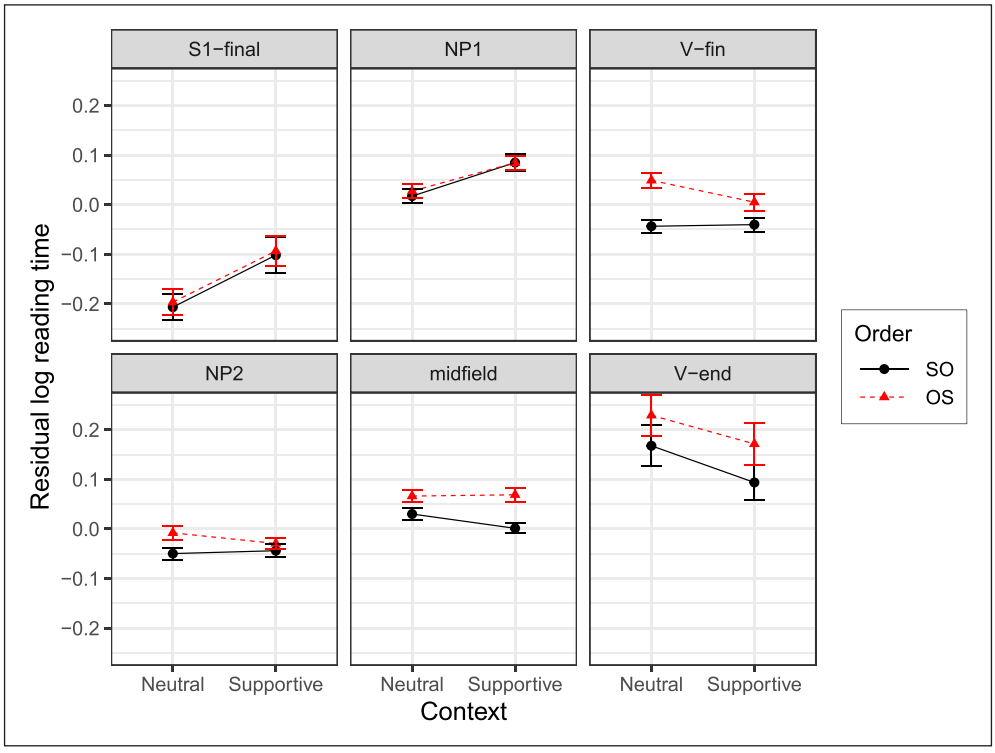


Figure 2: Mean reading times for the last word of the context sentence and the different regions of the target sentence in Experiment 1. Error bars show 95% confidence intervals.

Reading times for the finite verb (region V-fin) were higher for OS than for SO sentences, resulting in a significant effect of Order, whereas Context was not significant and the interaction between Order and Context was marginally significant. On the NP following the finite verb (region NP2), OS sentences were again read slower than SO sentences, resulting in a significant

effect of Order, but neither Context nor the interaction of Context and Order were significant. When V-fin and NP2 are combined to a single region, which can be considered the spill-over region of the sentence-initial NP, Order and the interaction of Order and Context are significant. This interaction reflects the finding that context had no effect on SO sentences in the joint V-fin/NP2 region (–0.048 vs. –0.044; $t = 0.89$, n.s.), but on OS sentences, which were read faster in this region when preceded by a supportive context (–0.012 vs. 0.022; $t = 3.53$, $p < 0.01$).

Table 2: Mixed-effects models for residual log reading times in Experiment 1.

Region	Contrast	Estimate	SE	df	t value	p value
S1-final	Order	0.008	0.022	1325.2	0.37	n.s.
	Context	0.109	0.022	1324.7	4.93	< 0.01
	Order × Context	–0.007	0.044	1325.4	–0.16	n.s.
NP1	Order	0.004	0.012	1334.7	0.32	n.s.
	Context	0.062	0.012	1334.5	5.11	< 0.01
	Order × Context	–0.009	0.024	1334.8	–0.39	n.s.
V-fin	Order	0.069	0.013	1336.3	5.42	< 0.01
	Context	–0.020	0.013	1336.0	–1.58	n.s.
	Order × Context	–0.048	0.025	1336.4	–1.90	< 0.10
NP2	Order	0.031	0.011	1335.1	2.77	< 0.01
	Context	–0.007	0.011	1334.8	–0.67	n.s.
	Order × Context	–0.026	0.022	1335.5	–1.16	n.s.
V-fin&NP2	Order	0.049	0.009	1334.5	5.24	< 0.01
	Context	–0.014	0.009	1334.3	–1.52	n.s.
	Order × Context	–0.038	0.019	1334.8	–2.02	< 0.05
Midfield	Order	0.051	0.009	1329.2	5.57	< 0.01
	Context	–0.013	0.009	1328.9	–1.40	n.s.
	Order × Context	0.032	0.018	1329.1	1.74	< 0.1
V-end	Order	0.072	0.021	1282.6	3.52	< 0.01
	Context	–0.066	0.021	1281.9	–3.19	< 0.01
	Order × Context	0.013	0.041	1282.2	0.31	n.s.

In the region following the second NP, which contains all words after NP2 except for the clause-final verb (region Midfield), Order was significant, the interaction of Order and Context

was marginally significant whereas Context was not. In this region, the interaction reflects the finding of a context effect for SO sentences, which were read faster in supportive contexts (–0.003 vs. 0.030; $t = 2.52$, $p < 0.05$), but no context effect for OS sentences (0.062 vs. 0.065; $t = 0.28$, n.s.). The sentence-final verb (region V-end) was read faster in supportive contexts than in neutral contexts, resulting in a main effect of Context. In addition, the final verb was read faster in SO sentences than in OS sentences, resulting in a main effect of Order. The interaction of Order and Context was not significant.

2.2.2 Comprehension questions

Figure 3 shows the accuracy of answering the comprehension questions as well as the mean reaction times for correct answers. The corresponding mixed-effects models are summarized in **Tables 3** and **4**. The accuracy data show significant main effects of Order and Question Type as well as a significant interaction between Order and Question Type. The factor Context was neither significant as a main factor nor involved in any significant interaction. As can be seen in **Figure 3**, accuracy was high for SO sentences followed by an agent question but low when followed by a patient question. Accuracy for OS sentences was also low, but still somewhat higher than for SO sentences followed by a patient question. Pairwise comparisons revealed the following ranking between the four combinations of order and question type: Accuracy was higher for SO order/agent questions than for OS order/agent questions (89% vs. 69%; $z = 6.77$, $p < 0.01$); OS order/agent questions resulted in a higher accuracy than OS order/patient questions, but the difference was not significant (69% vs. 62%; $z = 0.77$, $p > 0.1$); the accuracy for OS order/patient questions, finally, was higher than the accuracy for SO order/patient questions (62% vs. 52%; $z = 2.96$, $p < 0.01$).

Reaction times to correctly answered questions were longer for patient questions than for agent questions (3810 ms vs. 3064 ms), resulting in a main effect of Question Type, as shown in **Table 4**. Furthermore, the effect of Order was significant, but this effect is qualified by a significant interaction between Order and Question Type. For agent questions, reaction times were significantly longer after OS sentences than after SO sentences (3283 ms vs. 2901 ms; $t = 5.26$, $p < 0.01$). For patient questions, reaction times were longer for SO sentences than for OS sentences, but this difference did not reach significance (3889 ms vs. 3744 ms; $t = 1.08$, $p > 0.1$). In addition, the interaction between Context and Question Type was significant. This reflects the finding of longer reaction times for agent questions in neutral than in supportive contexts whereas the reverse holds for patient questions. However, the difference between reaction times in neutral versus supportive contexts was not significant for agent questions (3136 ms vs. 2993 ms; $t = 1.11$, $p > 0.1$) and only marginally significant for patient questions (3879 ms vs. 3745 ms; $t = 1.89$, $p = 0.059$).

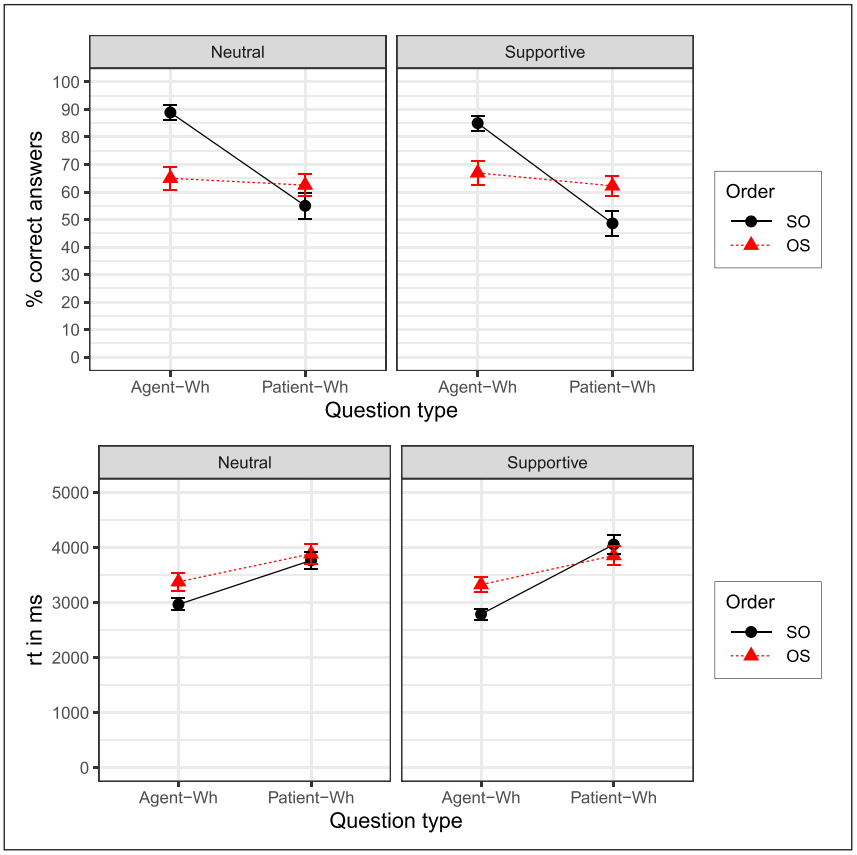


Figure 3: Percentages of correct answers to the comprehension questions and reaction times for correctly answered questions in Experiment 1. Error bars show 95% confidence intervals.

Table 3: Mixed-effects model for question accuracy in Experiment 1.

Contrast	**Estimate**	**SE**	**z value**	**p value**
Formula: *correct ~ QuestionType * Order * Context + (1 + QuestionType || participant) + (1 | sentence)*				
Intercept	0.921	0.149	6.20	
Order	0.485	0.134	3.61	< 0.001
Context	0.131	0.131	0.99	n.s.
QuestionType	1.192	0.277	4.31	< 0.001
Order × Context	0.376	0.263	1.43	n.s.
Order × QuestionType	1.933	0.269	7.18	< 0.001
Context × QuestionType	–0.003	0.263	–0.01	n.s.
Order × Context × QuestionType	0.230	0.526	0.44	n.s.

Table 4: Mixed-effects model for reaction times for comprehension questions in Experiment 1.

Contrast	Estimate	SE	df	t value	p value
Formula: *log(RT question) ~ QuestionType * Order * Context + (1 | participant) + (1 | sentence)*					
Intercept	8.085	0.036	50.7	227.18	
Order	–0.053	0.020	836.6	–2.61	< .01
Context	–0.014	0.020	829.4	–0.71	n.s.
QuestionType	–0.185	0.049	16.6	–3.81	< .01
Order × Context	–0.031	0.040	835.0	–0.76	n.s.
Order × QuestionType	–0.172	0.041	839.8	–4.24	< .01
Context × QuestionType	0.086	0.040	829.5	2.16	< .05
Order × Context × QuestionType	0.098	0.081	835.5	1.22	n.s.

2.2.3 Relationship between reading times and question accuracy

According to the GE model, misinterpretation errors observed for noncanonical sentences are the result of applying heuristics like the agent-before-patient heuristic, which are fast but error prone. Analyzing sentences with noncanonical word order using algorithmic procedures takes more time but typically delivers the correct interpretation. As argued above, the GE model therefore predicts that reading times for noncanonical sentences should be longer when the following question was answered correctly than when the question was answered incorrectly. Sentences with canonical word order receive the correct interpretation whether analyzed by heuristics or algorithmic means. For them, reading times should not differ depending on whether the following question was answered correctly or not.

In order to test the predicted relationship between reading times and correctness, **Figure 4** shows mean raw reading times per word broken down by word order of the target sentence and correctness of the following wh-question. To test for significant effects, a linear mixed-effects model was fitted to the data, with logarithmic raw reading times as dependent variable, Order and Correctness as fixed effects, and participants and items as random effects. In accordance with the analysis presented above, reading times were longer for OS than for SO sentences, resulting in a main effect of Order ($\beta = 0.049$; SE $= 0.011$; $t = 4.60$, $p < 0.01$). Sentences followed by a correct answer were read somewhat faster than sentences followed by an incorrect answer, resulting in a main effect of Correctness ($\beta = 0.027$; SE $= 0.012$; $t = 2.34$, $p < 0.05$). The interaction between Order and Correctness was not significant ($\beta = –0.0037$; SE $= 0.022$; $t =$ –0.15, n.s.).

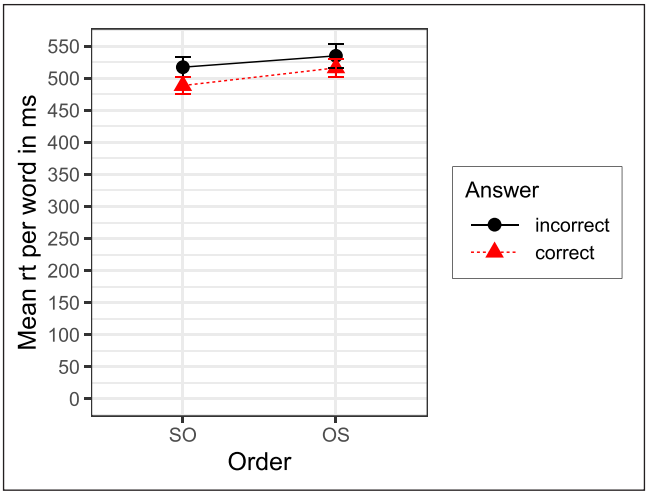


Figure 4: Mean reading time per word for trials with correctly and incorrectly answered questions in Experiment 1.

2.2.4 Post-hoc analysis: Acceptability and choice preference of SO and OS sentences in context

We constructed the neutral and the supportive contexts of Experiment 1 based on our understanding of the relevant literature on word order variation in German. However, the finding that supportive contexts only had a weak facilitative effect on the processing of noncanonical sentences raises the question of whether the force of the supportive context to license OS order was too weak. To address this issue, we ran a post-hoc norming study after the main experiment was completed. Two different tasks were used to assess the force of the context to license OS order: acceptability ratings and a forced-choice continuation study (see Bresnan & Ford, 2010; Rosenbach, 2005). The tests were run on IBEX farm (Drummond et al., 2016) with different sets of participants.

For the rating study, context and target sentence were presented together on a single screen. Participants were asked to rate how well the second sentence fits the first sentence on a scale from 1 ("Does not fit at all") to 7 ("Fits perfectly"). 18 students of Goethe University Frankfurt received course credit for participation in the test. The 24 experimental items were mixed with 42 filler items.

The mean acceptability values for the materials of Experiment 1 are shown on the left side in **Figure 5**. Ordinal mixed-effects models were fitted to participants' responses using the ordinal package (Christensen, 2019). Results show significantly higher acceptability ratings for SO compared to OS sentences ($z = 5.852, p < 0.01$), and significantly higher ratings for sentences in supportive contexts than for sentences in neutral contexts ($z = 7.410, p < 0.01$). The interaction

was significant as well ($z = 2.937$, $p < 0.01$), reflecting the finding that the acceptability of SO sentences improved in supportive contexts to an even greater extent than the acceptability of OS sentences.

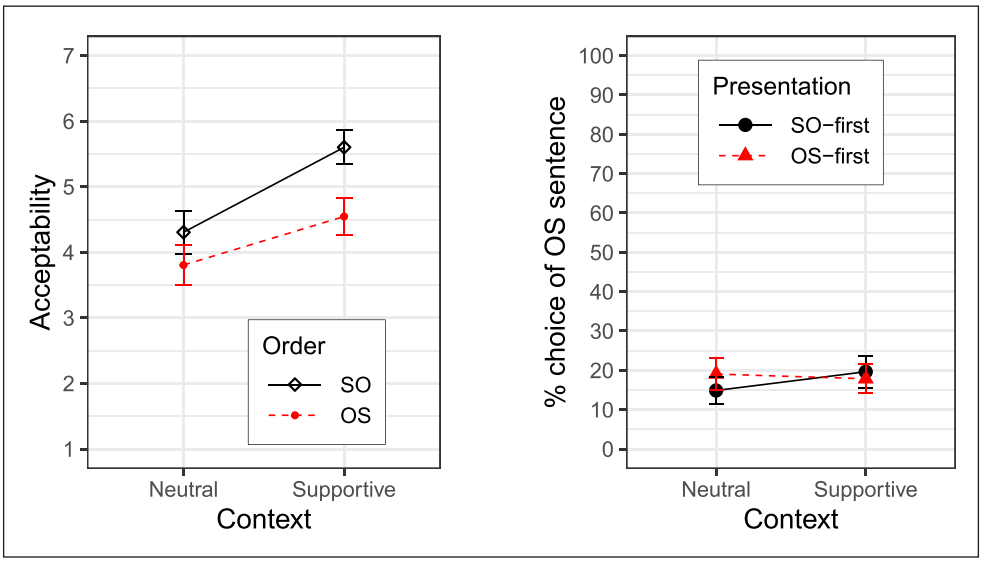


Figure 5: Results of the post-hoc tests run to assess the force of a supportive context to license OS order in Experiment 1: Mean acceptability ratings in the acceptability rating test (left) and percentages of how often the OS continuation was chosen in the forced-choice continuation test (right).

In the forced-choice continuation study, a context sentence was always displayed together with two versions of the target sentence: one with SO order and one with OS order. Participants had to select the version of the target sentence which they would prefer as continuation of the context sentence. The order of presenting the two versions of the target sentence was a second factor (Presentation: SO-first vs. OS-first). 28 students from Goethe University Frankfurt received course credit for participation. The 24 experimental items were mixed with 36 items serving as fillers.

The percentages of selecting the OS target sentence as preferred continuation are shown on the right side of **Figure 5**. Overall, the OS target sentence was selected by participants in 18% of all cases. A logistic mixed-effects model revealed that neither the context nor the order of the two alternative continuations on the display (factor Presentation) had any significant effect. The interaction was not significant either. In sum, the results suggest that the supportive context condition improves acceptability of OS sentences as intended. However, the acceptability of SO sentences increased to an even stronger degree, and the strong preference for an SO continuation persisted. How these findings relate to the reading time results is considered in the discussion section, to which we turn now.

2.3 Discussion

In line with previous studies, Experiment 1 observed increased online processing difficulty for noncanonical OS sentences. With the exception of the sentence-initial NP, reading times for OS sentences were higher than reading times for SO sentences at all positions. The OS disadvantage was particularly robust at the sentence-final verb. The context manipulation affected reading times in several regions. On the sentence-initial NP, sentences following a supportive context elicited higher reading times than sentences following a neutral context. This effect is surprising, given that a supportive context established the sentence-initial NP as given, and evidence from prior research points to a processing advantage for given NPs compared to new NPs (Kaiser & Trueswell, 2004; Koizumi & Imamura, 2017). However, a similar effect was already visible on the last word of the preceding context sentence, possibly reflecting higher sentence wrap-up efforts due to the need to set up new discourse referents (Just & Carpenter, 1980). We therefore consider the effect of context on the sentence-initial NP a spillover effect.

On the finite auxiliary verb and the following NP2, context had no effect on reading times for SO sentences, but for OS sentences reading times were faster in supportive contexts than in neutral contexts, although the interaction was significant only when both regions were analyzed together. In the region following NP2, the context effect was reversed, with SO sentences now benefiting from a supportive context whereas OS sentences did not. On the sentence-final verb, a supportive context decreased reading times for SO and OS sentences to similar degrees. The facilitative effect of supportive contexts for both SO and OS sentences mirrors the results from the post-hoc acceptability test, which revealed higher acceptability values in supportive contexts for both SO and OS sentences.

In contrast to reading times as a measure of online processing, context had no main effect on how often comprehension questions were answered correctly, nor was it involved in any interaction. Order and Question Type, in contrast, strongly affected performance on comprehension questions, showing a robust interaction. For SO sentences, accuracy was much higher following agent questions than following patient questions. For OS sentences, accuracy was generally low, with no significant difference between agent and patient questions. For reaction times to comprehension questions, a similar pattern emerged. Longer reaction times were elicited in response to agent questions following OS compared to SO sentences. On the other hand, no reliable difference in reaction times was observed for patient wh-questions, for which reaction times were equally high following SO and OS target sentences.

The analysis of reading times depending on correctness on comprehension questions revealed increased reading times for OS sentences regardless of whether the comprehension question was answered correctly or not. We also found an effect of correctness. Sentences that were answered correctly were read faster than sentences that were answered incorrectly. The interaction between order and correctness was not significant.

In sum, the results of Experiment 1 show a robust effect of word order on online processing. The context manipulation had some moderate effects on the reading times but did not affect the comprehension accuracy for either SO or OS sentences. To follow up on the results obtained in Experiment 1, Experiment 2 tests whether the online and offline result patterns change if discourse properties provide stronger cues in favor of OS order.

3. Experiment 2

Recent investigations of the conditions favoring the production of OS sentences in German show that besides the referents' discourse status (e.g., given or new), the particular referential expression used for the referent in object position is a main determinant of whether to front the object or not. A corpus study of German Wikipedia texts reported in Bader and Portele (2019) found 18% OS order for sentences with definite objects, but 76% OS order for sentences with demonstrative objects. Corroborating acceptability results have been provided by Bader and Portele (2021) in an investigation of short texts as in (5), in which the object of the target sentence was a definite or a demonstrative NP (*den Kollegen* versus *diesen Kollegen*).

(5) Context: Ich habe gestern einen ehemaligen Kollegen getroffen.
'I met a former colleague yesterday.'

a. SO target: Ich habe (den/diesen) Kollegen sofort wiedererkannt.
I.NOM have the/this.ACC colleague immediately recognized
'I recognized him immediately.'

b. OS target: (Den/Diesen) Kollegen habe ich sofort wiedererkannt.
the/this.ACC colleague have I.NOM immediately recognized
'Him, I recognized immediately.'

Bader and Portele (2021) found that OS sentences received somewhat lower acceptability ratings than SO sentences when the object was a definite NP, whereas OS sentences were as acceptable as SO sentences when the object was a demonstrative NP.

The main goal of Experiment 2 was to test whether the higher rate of object-initial sentences with demonstrative objects is reflected in online processing and offline comprehension. To this end, Experiment 2 manipulated the order of subject and object in the target sentence (SO vs. OS) as well as the referential form of the object NP (demonstrative vs. definite NP). This is illustrated in (6).

(6) Context: Heute hat es viel Unruhe vor und während des Unterrichts gegeben. Der neue Lehrer war mal wieder nicht zufrieden mit einem Schüler aus der fünften Klasse.
'Today there was a lot of unrest before and during class. Once again, the new teacher was not satisfied with a student from the 5th grade class.

a. SO target: Der Lehrer hat den/diesen Schüler deshalb angeschrien.
the teacher has the/this student therefore yelled-at
'The teacher therefore yelled at the/this student.'

b. OS target: Den/Diesen Schüler hat der Lehrer deshalb angeschrien.
the/this student hat the teacher therefore yelled-at
'The teacher therefore yelled at the/this student.'

Target sentences were preceded by a supportive context combining the neutral and supportive context sentences from Experiment 1. Therefore, the context licenses OS order in all conditions but is also compatible with SO order. For each target sentence, the context was identical across all conditions.

In addition to effects of referential form, Experiment 2 was intended to test whether the strong effect of question type observed in Experiment 1 replicates.

3.1 Method

3.1.1 Participants

Due to the COVID-19 pandemic, all parts of Experiment 2 had to be run on the Internet, again using IBEX farm. The main experiment tested 60 native speakers of German who were recruited over Prolific (http://prolific.co), with filter criteria "native language = German" and "Desktop Only." In addition, participants were asked to confirm at the beginning of the session that their native language was German. Participants were naive with respect to the purpose of the experiment. Their mean age was 29.9, ranging from 18 to 63. Participants received £3.75 for participation.

3.1.2 Materials

Experiment 2 used the 24 active declarative sentences constructed for Experiment 1, with each sentence occurring in four versions (see **Table 5**). As before, the factor Order varied the order of subject and object (SO vs. OS). The new factor Object manipulated the determiner of the object NP, which was either a definite article or a demonstrative determiner.

Each target sentence was preceded by a two-sentence context. The first context sentence (C1) was the context sentence from the neutral context condition of Experiment 1. For the second context sentence (C2), we used the context sentence from the supportive context condition of Experiment 1. Context sentences and target sentence were immediately followed by the same wh-questions as in Experiment 1. The factor Question Type (agent-wh vs. patient-wh) was again varied between sentences: half of the experimental sentences were followed by an agent question, the other half by a patient question.

Table 5: A complete stimulus item for Experiment 2 including a patient-wh question

Context:	[C1]	Wie jedes Jahr hat auch dieses Jahr wieder ein rauschendes Fest im Schloss stattgefunden.
		'Like every year, a wonderful celebration took place at the castle this year.'
	[C2]	Der betagte König hat gleich zu Beginn einen Botschafter in einem ausgefallenen Kostüm begrüßt.
		'Already at the beginning, the old king greeted an ambassador in a fancy masquerade costume."
Target:	SO	Der König hat den/diesen Botschafter dabei trotz der Maske erkannt.
		'The king recognized the/this ambassador despite the mask.'
	OS	Den/Diesen Botschafter hat der König dabei trotz der Maske erkannt.
		'The king recognized the/this ambassador despite the mask.'
Question:		Wen hat jemand erkannt? – König – Botschafter
		'Who recognized someone? – King – Ambassador'

3.1.3 Pre-test norming study

To confirm that our manipulation facilitates OS order as intended, we conducted an acceptability rating study and a forced-choice continuation study on the materials, this time before running the actual experiment. The same procedures were used as in the post-hoc tests of Experiment 1. Participants were recruited for course credit from the pool of students at Goethe University Frankfurt.

For the acceptability rating study, 19 participants read the two context sentences followed by the respective target sentence in one of the four conditions resulting from the factors Order (SO vs. OS) and Object (definite vs. demonstrative NP). Participants were asked to rate on a 7-point scale how well the target sentence fits the preceding context. The 24 experimental items were combined with 42 filler items. The mean acceptability values are shown on the left side of **Figure 6**. Ordinal mixed-effects models fitted to the response data revealed higher rating for SO compared to OS sentences (5.4 vs. 5.1; $z = 2.92$, $p < 0.01$), but no effect of Object (5.1 vs. 5.4; $z = 1.42$, $p = 0.15$). The interaction between Order and Object failed to reach significance ($z = 1.53$, $p = 0.13$). Pairwise comparisons showed that with demonstrative objects, ratings for OS did not differ from ratings for SO (5.4 vs. 5.3, $z = 1.02$, $p = 0.31$) whereas OS sentences were rated significantly worse than SO sentences for definite objects (5.4 vs. 4.9, $z = 3.10$, $p < 0.01$).

For the forced-choice continuation task, we presented the two context sentences followed by the SO and the OS version of the respective target sentence. 33 participants read each context and

then clicked on the target sentence version they preferred as continuation. The object of the target sentence had either a definite or a demonstrative determiner. Either the SO version was presented above the OS version or the other way around. The 24 experimental items were interspersed with 36 fillers. As shown in **Figure 6**, there was again an overall preference to select the SO variant of the target sentences, but using demonstrative objects significantly increased the percentage of OS continuations from 19% to 38% ($z = 6.42, p < 0.01$). The order of presentation and the interaction between determiner and presentation order were not significant (presentation order: $z = 1.38, p = 0.17$; interaction: $z = –1.23, p = 0.22$). Taken together, the pretests show that the combined force of a supportive context and a demonstrative object NP substantially facilitates OS order.

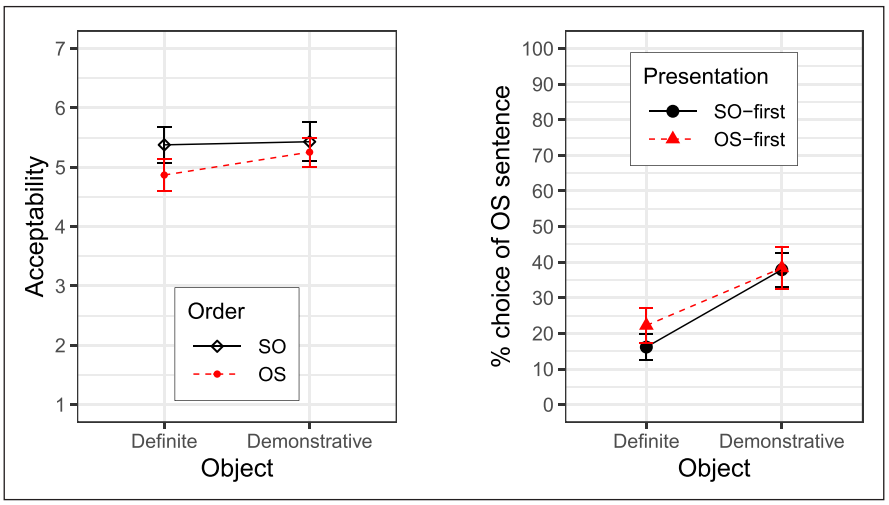


Figure 6: Results of the pre-tests run to assess the joint force of context and determiner to license OS order in Experiment 2: Mean acceptability ratings in the acceptability rating test (left) and percentages of how often the OS continuation was chosen in the forced-choice continuation test (right).

3.1.4 Procedure

Due to technical limitations resulting from the switch to the Internet, no reading span data were collected in Experiment 2 and participants engaged in the selfpaced reading task only. At the beginning of the session, participants read the instruction page and completed a simple form collecting demographic data including age and L1. Participants were also asked to give informed consent to participate in the study. The selfpaced reading procedure mirrored the procedure of Experiment 1 as closely as possible. The three sentences of each trial appeared together on a single screen. Sentences were presented using a non-cumulative moving window paradigm. Participants progressed through the sentences word by word using the space bar. Only the active word was visible and all other words were masked by underlines.

Immediately after the last word of the target sentence, the comprehension question was displayed along with two answer options referring either to the agent or the patient of the target sentence. Answer options were marked with “1” and “2”. Participants selected an answer option either by performing a mouse click on the respective option or by selecting “1” or “2” on the keyboard. Participants did not get feedback concerning the correctness of their answer. After participants had selected an answer, the next trial was initiated. Before starting the selfpaced reading task, participants received three practice trials.

3.2 Results

One participant was excluded from the analysis because of only 63% accuracy for the questions following the filler sentences (range of remaining 59 participants: 80%–100%, mean = 97.6%).

3.2.1 Reading times

Logarithmic residual reading times were computed in the same way as for Experiment 1. **Figure 7** and **Figure 8** show the time course of reading the final word of the context sentence and the different regions of the target sentence. The results of the linear mixed-effects models for the individual regions are summarized in **Table 6**.

For the final word of the context sentence, no significant effects were observed, which was expected given that the context was identical across conditions. On the sentence-initial NP (region NP1), the effect of Order was marginally significant, the effect of Object was not significant, but the interaction was significant. The interaction reflects the finding that for sentences with a definite object, reading times on NP1 were significantly longer for OS than for SO sentences (0.083 vs. 0.033; $t = 3.01$, $p < 0.01$), whereas no significant reading time difference depending on word order showed up for sentences with a demonstrative object (0.048 vs. 0.043; $t = 0.28$, n.s.). For all remaining regions, the factor Order showed a significant effect whereas neither the factor Object alone nor the interaction between Order and Object was significant.

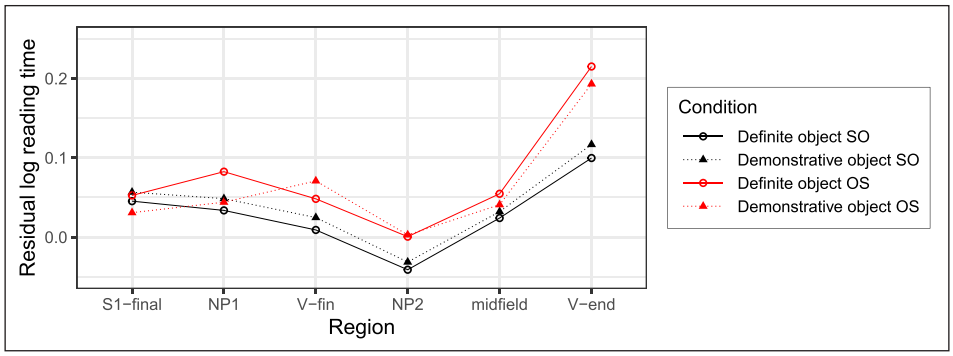


Figure 7: Mean reading times for the last word of the context sentence and the different regions of the target sentence in Experiment 2.

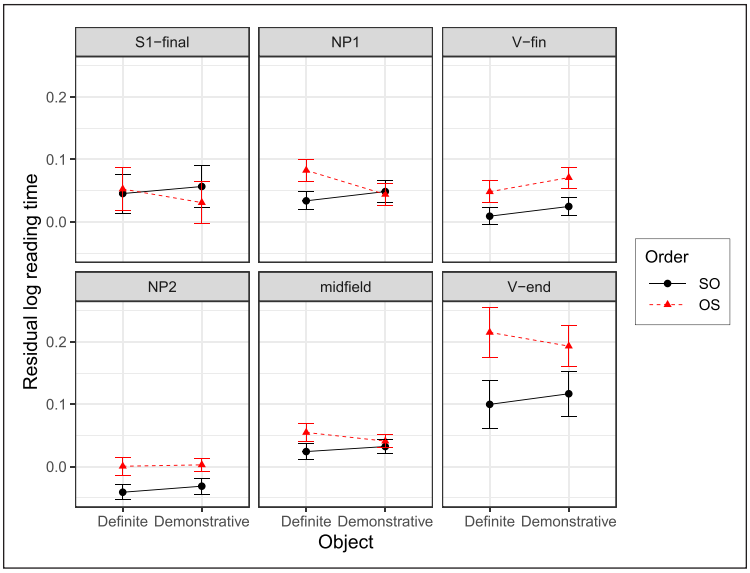


Figure 8: Mean reading times for the last word of the context sentence and the different regions of the target sentence in Experiment 2. Error bars show 95% confidence intervals.

Table 6: Mixed-effects models for residual log reading times in Experiment 2.

Region	Contrast	Estimate	SE	df	t value	p value
S1-final	Order	–0.008	0.018	1302.3	–0.43	n.s.
	Object	–0.005	0.018	1302.6	–0.27	n.s.
	Order × Object	–0.034	0.037	1302.3	–0.92	n.s.
NP1	Order	0.022	0.012	1318.7	1.93	< 0.10
	Object	–0.012	0.012	1318.9	–1.02	n.s.
	Order × Object	–0.053	0.023	1318.8	–2.32	< 0.05
V-fin	Order	0.043	0.013	1332.8	3.27	< 0.01
	Object	0.019	0.013	1332.9	1.47	n.s.
	Order × Object	0.007	0.026	1332.8	0.27	n.s.
NP2	Order	0.038	0.010	1327.9	3.83	< 0.01
	Object	0.006	0.010	1328.2	0.58	n.s.
	Order × Object	–0.007	0.020	1328.0	–0.35	n.s.
Midfield	Order	0.021	0.009	1317.0	2.45	< 0.05
	Object	–0.000	0.009	1315.4	–0.03	n.s.
	Order × Object	–0.028	0.017	1315.1	–1.64	n.s.
V-end	Order	0.096	0.021	1300.6	4.64	< 0.01
	Object	–0.001	0.021	1300.6	–0.04	n.s.
	Order × Object	–0.034	0.041	1300.7	–0.81	n.s.

3.2.2 Comprehension questions

Accuracy and reaction times for answering questions in Experiment 2 were analyzed in the same way as in Experiment 1. As revealed by **Figure 9**, the percentages of correct answers and the reaction times for correct answers show the same patterns as in Experiment 1. The generalized mixed-effects model for the accuracy results, summarized in **Table 7**, again revealed significant main effects of Order and Question Type as well as a significant interaction between them. Accuracy was higher for SO order/agent question than for OS order/agent question (91% vs. 78%; $z = 4.880$, $p < 0.01$). Accuracy for OS order/agent question in turn was higher than accuracy for OS order/patient question (78% vs. 65%; $z = 3.891$, $p < 0.01$). Finally, OS order/patient question resulted in a higher accuracy than SO order/patient question (65% vs. 53%; $z = 3.645$, $p < 0.01$).

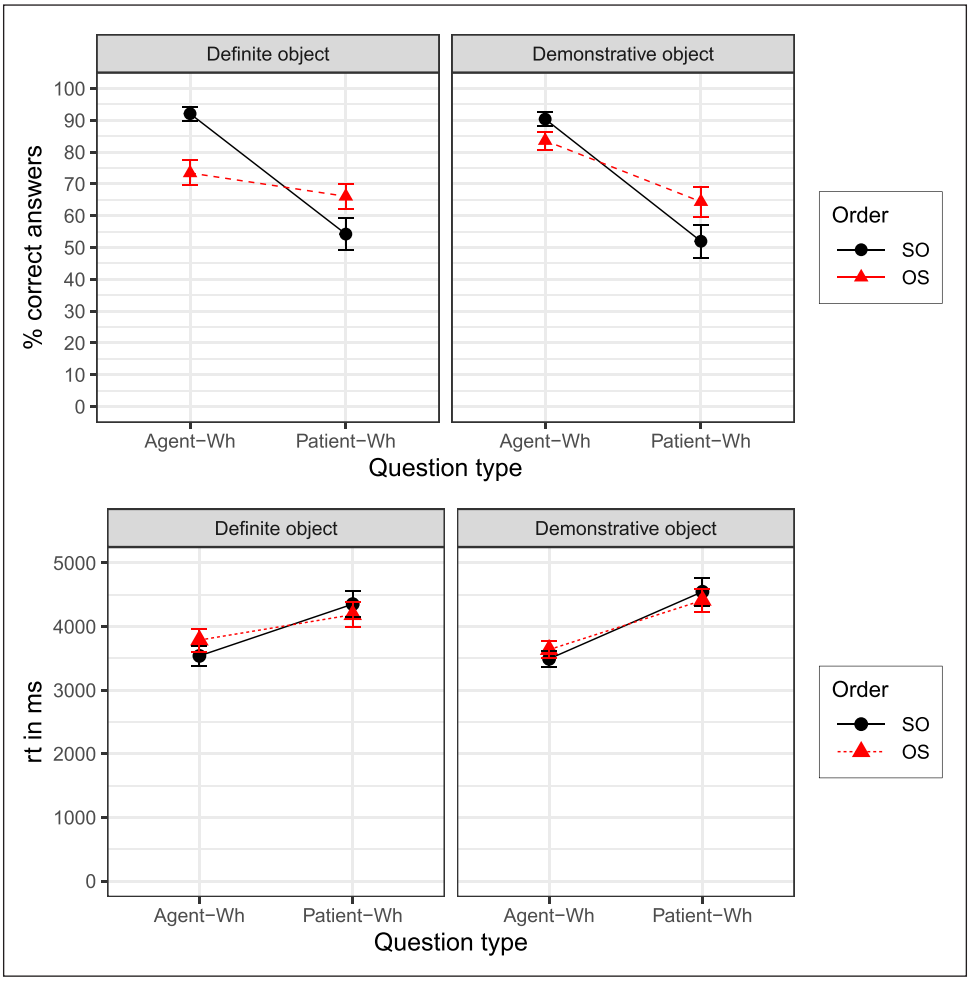


Figure 9: Percentages of correct answers to the comprehension questions and reaction times for correctly answered questions in Experiment 2. Error bars show 95% confidence intervals.

Table 7: Mixed-effects model for question accuracy in Experiment 2.

Contrast	Estimate	SE	z value	p value
Formula: *correct ~ QuestionType * Order * Object + (1 | participant) + (1 | sentence)*				
Intercept	1.340	0.165	8.12	
Order	–0.257	0.146	–1.75	< 0.1
Object	0.068	0.146	0.47	n.s.
QuestionType	–1.693	0.173	–9.79	< 0.001
Order × Object	0.475	0.293	1.62	n.s.
Order × QuestionType	1.759	0.295	5.96	< 0.001
Object × QuestionType	–0.339	0.293	–1.16	n.s.
Order × Object × QuestionType	–0.900	0.586	–1.54	n.s.

The factor Object was not significant, nor was any interaction involving Object. An inspection of **Figure 9** shows that accuracy was almost identical for sentences with definite NP and sentences with demonstrative NP in three combinations of the factors Order and Question Type (difference $< 2\%$), but in the condition OS order followed by an agent wh-question, accuracy was about 11% higher for sentences with demonstrative NP than for sentences with definite NP (84% versus 73%). If reliable, this should have resulted in a three-way interaction, but with $p = 0.12$ the three-way interaction failed to reach significance. Since getting a significant three-way interaction is difficult when it is carried by a single, relatively small difference, we also computed two-way analyses separately for SO and OS sentences. For SO sentences, there was only a significant main effect of Question Type ($\beta = 2.748$; SE $= 0.257$; $z = 10.68$, $p < 0.01$). For OS sentences, in contrast, there was a significant main effect of Question type ($\beta = 0.763$; SE $= 0.202$; $z = 3.79$, $p < 0.01$) and also a significant interaction between Object and Question Type ($\beta = 0.749$; SE $= 0.361$; $z = 2.07$, $p < 0.05$). We therefore tentatively conclude that answering agent questions following an OS sentence is eased when the fronted object is a demonstrative NP.

Reaction times to correct answers are also shown in **Figure 9**. Although a pattern similar to Experiment 1 is visible, the corresponding linear mixed-effects model for the logarithmic reaction times, summarized in **Table 8**, reveals only a main effect of Question Type. Agent questions were answered significantly faster than patient questions (3576 ms vs. 4431 ms).

3.2.3 Relationship between reading times and question accuracy

As for Experiment 1, we analyzed the relationship between reading times, taken as mean raw reading time per word, and the correctness of the answer selected for the wh-question

following each target sentence. As shown in **Figure 10**, reading times were higher for OS than for SO sentences, in line with the reading time analysis presented above. **Figure 10** also shows small differences between sentences followed by correctly and incorrectly answered questions. However, while the factor Order was significant ($\beta = 0.047$; SE $= 0.010$; $t = 4.80$, $p < 0.01$), neither the factor Correctness ($\beta = 0.013$; SE $= 0.011$; $t = 1.18$, n.s.) nor the interaction between Order and Correctness ($\beta = 0.028$; SE $= 0.020$; $t = 1.41$, n.s.) was significant.

Table 8: Mixed-effects model for reaction times for comprehension questions in Experiment 2.

Contrast	Estimate	SE	df	t value	p value
Formula: *log(RT question) ~ QuestionType * Order * Object + (1 | participant) + (1 | sentence)*					
Intercept	8.192	0.035	65.1	232.64	
Order	0.029	0.021	907.1	1.36	n.s.
Object	0.027	0.021	904.8	1.31	n.s.
QuestionType	0.161	0.033	25.0	4.95	< 0.001
Order × Object	0.001	0.043	905.4	0.02	n.s.
Order × QuestionType	–0.056	0.043	909.6	–1.31	n.s.
Object × QuestionType	0.050	0.043	905.8	1.17	n.s.
Order × Object × QuestionType	0.020	0.086	905.6	0.23	n.s.

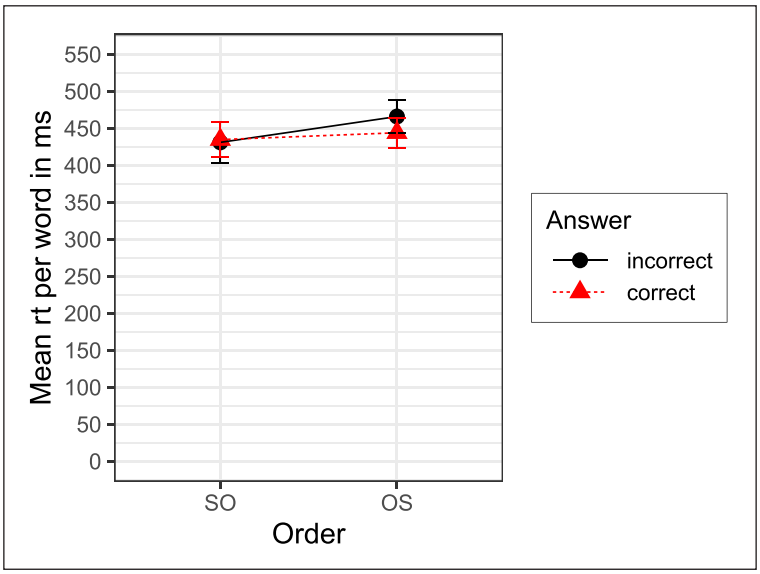


Figure 10: Mean reading time per word for trials with correctly and incorrectly answered questions in Experiment 2.

3.3 Discussion

Consistent with Experiment 1, Experiment 2 found that OS sentences elicited longer reading times compared to SO sentences. For sentences with a definite object, this effect surfaced in all regions starting with the initial NP, and it was again particularly robust on the sentence-final verb. For sentences with a demonstrative object, the reading time penalty for OS order started at the finite verb. On the sentence-initial NP, however, reading times for OS sentences (initial NP = demonstrative object) and SO sentences (initial NP = subject) did not differ. The online data for the sentence-initial NP thus mirror prior corpus and experimental findings, including the pretest results of Experiment 2, that demonstrative objects are particularly favorable for using OS order. The online data for the remaining part of the sentences, on the other hand, show a disadvantage for OS sentences in comparison to SO sentences, independent of the particular determiner used for the object.

The accuracy results look very similar to those of Experiment 1. Experiment 2 again found a robust interaction between Order and Question Type. For SO sentences, accuracy was much higher for agent questions than for patient questions, whereas the accuracy for OS questions was in between. In addition, the object manipulation led to a slight improvement of accuracy for OS sentences followed by agent questions. As in Experiment 1, agent questions were responded to substantially faster than patient questions.

In sum, Experiment 2 provides further evidence that discourse-related properties licensing noncanonical order alleviate online processing difficulty for OS sentences, although restricted to the sentence-initial NP. Comprehension accuracy improved marginally at best for OS sentences when OS order was favored by discourse-related properties. Instead, accuracy was, as in Experiment 1, jointly determined by word order of the target sentence and question type.

4. Experiment 3

A major finding of the two preceding experiments has been that performance on comprehension questions varied strongly depending on the particular combination of target sentence and comprehension question. Accuracy was highest for SO sentences followed by an agent question, but the poorest accuracy was also found for SO sentences, namely when followed by a patient question. For OS sentences, accuracy was reduced throughout, but agent questions were again answered better than patient questions. The observed interaction makes it difficult to assess comprehension accuracy.

Given that in prior research, sentences with canonical word order were found to be least susceptible to misinterpretation errors, the high number of incorrect answers for SO sentences followed by a patient question seems particularly surprising. But even for OS sentences, patient questions caused more errors. However, patient questions in Experiments 1 and 2 were always OS sentences, and therefore sentences with noncanonical order themselves. It may thus well be that even if participants understood the target sentence correctly, they misinterpreted the

following patient question and therefore gave an incorrect answer. This would be the case, for example, if participants used the NVN ('Agent Verb Patient') heuristic to analyze both a target sentence showing SO order, and the following patient question showing OS order.

The aim of Experiment 3 was to test whether patient questions are more difficult to answer even if patient questions no longer show OS order. To this end, Experiment 3 required participants to answer one-argument questions as illustrated in (7).

(7)

SO target:	Der Stürmer hat diesen Verteidiger dann auch ziemlich rüde gefoult.
	'The striker then fouled this defender very badly.'
OS target:	Diesen Verteidiger hat der Stürmer dann auch ziemlich rüde gefoult.
	'The striker then fouled this defender very badly.'
Agent question:	Wer hat gefoult? – Stürmer – Verteidiger
	'Who fouled? – Striker – Defender'
Patient question:	Wer wurde gefoult? / Stürmer – Verteidiger
	'Who was fouled? – Striker – Defender'

Intransitive active questions asked for the agent/subject of the target sentences. Simple passive questions (i.e., passive sentences without by-phrase) asked for the patient/object. If asking for the patient is inherently more difficult than asking for the agent, a similar result pattern should be found as before. On the other hand, since patient questions now also have a subject wh-phrase, potential difficulties arising from the OS order of patient questions should be eliminated.

Unlike the preceding experiments, question type was a within-item factor in Experiment 3. The only other factor was the order of subject and object in the target sentence. The object was always a demonstrative NP and the context was always supportive.

4.1 Method

4.1.1 Participants

Experiment 3 was again run using IBEX farm, with participants being recruited over Prolific. Experiment 3 tested 32 native speakers of German. As in Experiment 2, we used "native language = German" and "Desktop Only" as filtering criteria and asked participants to confirm at the beginning of the session that their native language was German. An additional criterion for recruitment was to exclude participants that had completed Experiment 2. Participants were naive with respect to the purpose of the experiment. Their mean age was 29.7, ranging from 19 to 52. Participants received £4.00 for participation.

4.1.2 Materials

Experiment 3 adapted the 24 experimental items from Experiment 2. There were two differences in comparison to Experiment 2 (see **Table 9**). First, the factor Object was dropped. All object

NPs were demonstrative NPs. Second, Question type was made a within factor, and the questions were made less complex by containing only a single argument. Thus, Experiment 3 varied the two factors Order (SO vs. OS) and Question Type (agent vs. patient wh-question), resulting in 4 versions for each experimental item. The question always used the same verb as the target sentence. For agent questions, the subject was replaced by a wh-phrase and the object was dropped. For patient questions, the target sentence was transformed to a passive sentence without a by-phrase; the subject of the resulting passive sentence was replaced by a wh-phrase. Since agent wh-questions did not contain an object, a few items from Experiment 2 were modified to make sure all target sentences contain optionally transitive verbs. If necessary, context sentences were modified accordingly to ensure optimal fit.

Table 9: A complete stimulus item for Experiment 3.

Context:	[C1]	Schon vor dem Spiel war der Ton ziemlich rau gewesen.
		‘Already before the game, the atmosphere had been rather charged.’
	[C2]	Der eingewechselte Stürmer hatte nämlich einen Verteidiger des Gegners mehrfach beleidigt.
		‘The new striker had insulted a defender of the opposing team several times.”
Target:	SO	Der Stürmer hat diesen Verteidiger dann auch ziemlich rüde gefoult.
		‘The striker then fouled this defender very badly.’
	OS	Diesen Verteidiger hat der Stürmer dann auch ziemlich rüde gefoult.
		‘This defender, the striker then fouled very badly.’
Question:	Agent	Wer hat gefoult? – Stürmer – Verteidiger
		‘Who fouled? – Striker – Defender’
Question:	Patient	Wer wurde gefoult? / Stürmer – Verteidiger
		‘Who was fouled? – Striker – Defender’

4.1.3 Procedure

Experiment 3 used the same selfpaced reading procedure as Experiment 2.

4.2 Results

Data preparation and statistical analysis proceeded in the same way as in the preceding experiments. The questions for the filler sentences were answered with a mean accuracy of 94.4%. All participants had an accuracy of at least 70% and therefore no one was excluded from the analysis.

4.2.1 Reading times

Figure 11 presents the reading times in each region broken down by the single factor Order. The results of the mixed-effects models for each region are summarized in **Table 10**. For the final word and the initial NP of the target sentence, the factor Order did not have a significant effect. For the finite verb, a marginally significant effect of Order is observed, with OS sentences leading to longer reading times than SO sentences. The next two regions, NP2 and Midfield, show again no significant effect of Order, but for the clause-final verb, reading times were significantly higher for OS than for SO sentences.

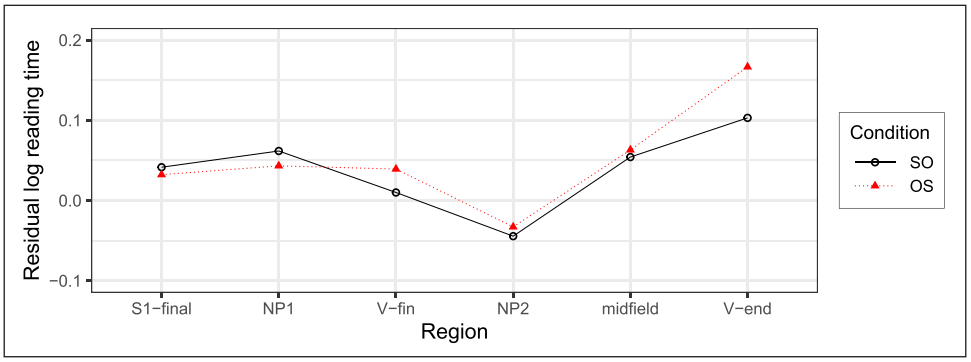


Figure 11: Mean reading times for the different regions of the target sentence in Experiment 3. Error bars show 95% confidence intervals.

Table 10: Mixed-effects models for residual log reading times in Experiment 3.

Region	Contrast	Estimate	SE	df	t value	p value
S1-final	Order	–0.011	0.030	700.7	–0.36	n.s.
NP1	Order	–0.017	0.017	703.3	–1.01	n.s.
V-fin	Order	0.029	0.016	709.2	1.79	< 0.10
NP2	Order	0.012	0.014	708.4	0.87	n.s.
Midfield	Order	0.009	0.012	705.3	0.76	n.s.
V-end	Order	0.064	0.028	701.9	2.31	< 0.05

4.2.2 Comprehension questions

Figure 12 shows the percentages of comprehension questions that were answered correctly in Experiment 3 and the reaction times for correct answers. The corresponding mixed-effects models are summarized in **Tables 11** and **12**. Overall, questions were answered correctly with an accuracy of 83%. As shown in **Figure 12**, accuracy was somewhat higher for SO than for OS sentences (86% versus 79%), resulting in a significant effect of Order. The factor Question Type

was not significant, nor the interaction between Order and Question Type. The mixed-effects model for the reaction times for correctly answered questions revealed significant effects of the two main factors and a significant interaction between them. While Question Type did not have a significant effect on SO sentences (2882 ms vs. 2839 ms; $t = 0.33$, $p > 0.1$), it affected OS sentences significantly, with agent questions needing about 500 ms longer for correct answers than patient questions (2801 ms vs. 3301 ms; $t = 3.25$, $p < 0.01$).

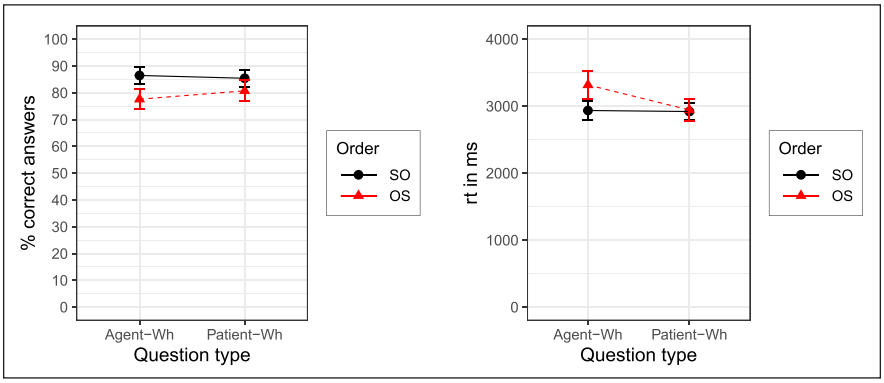


Figure 12: Percentages of correct answers to the comprehension questions and reaction times for correctly answered questions in Experiment 3. Error bars show 95% confidence intervals.

Table 11: Mixed-effects model for question accuracy in Experiment 3.

Contrast	**Estimate**	**SE**	**z value**	**p value**
Formula: *correct ~ QuestionType * Order + (1 | participant) + (1 | sentence)*				
Intercept	2.027	0.278	7.28	n.s.
Order	–0.588	0.209	–2.78	< 0.01
QuestionType	0.057	0.209	0.27	n.s.
Order × QuestionType	0.339	0.416	0.81	n.s.

Table 12: Mixed-effects model for reaction times for comprehension questions in Experiment 3.

Contrast	**Estimate**	**SE**	**df**	**t value**	**p value**
Formula: *log(RT question) ~ Order * QuestionType + (1 | participant) + (1 | sentence)*					
Intercept	7.933	0.043	41.4	183.50	
Order	0.057	0.027	563.8	2.13	< .05
QuestionType	–0.069	0.027	562.4	–2.58	< .05
Order × QuestionType	–0.114	0.053	561.2	–2.13	< .05

4.3 Discussion

Experiment 3 again found longer reading times for OS sentences compared to SO sentences. However, in contrast to the preceding experiments, the disadvantage for OS sentences was only significant on the finite verb and on the clause-final verb. On the sentence-initial NP1, reading times for SO sentences were slightly higher than reading times for OS sentences, but the difference was not significant. This is consistent with the absence of an effect of Order on NP1 in Experiment 2 for sentences with a demonstrative NP.

The main finding yielded by Experiment 3 is that presenting one-argument wh-questions instead of two-argument questions, as in Experiment 1 and 2, eliminated any effects of question type on the accuracy of answering comprehension questions, although an effect was still visible on the reaction times. Accuracy on comprehension questions was fairly high overall, both for agent and for patient questions. Note also that the accuracy values are close to the values obtained for SO sentences and active questions in the demonstrative object NP condition in Experiment 2.

5. General discussion

Noncanonical sentences have provided a long-standing challenge to models of the human parser. Across languages and sentence types, noncanonical sentences have been found to be more difficult to process than canonical sentences, as reflected in both online and offline measures. In order to shed further light on this finding, we ran three experiments investigating German active SO and OS sentences. Online processing was assessed using selfpaced reading. The final interpretation formed by comprehenders was probed by wh-questions. Two main questions guided our study: how does context affect the online processing and final interpretation of sentences with noncanonical word order, and what is the source of misinterpretation errors that have been reported for sentences with noncanonical word order. We discuss in turn how the results yielded by Experiments 1–3 bear on these questions.

5.1 Noncanonical sentences in context

In line with previous findings, our online results show that OS sentences are more difficult to process than SO sentences, and that this difficulty can be reduced by discourse properties licensing OS order. Facilitative effects showed up mainly at early sentence positions: on the finite verb and the following NP when the context was manipulated to license OS (Experiment 1), and on the sentence-initial NP when the referential form of the object NP favored OS order as well (Experiments 2 and 3). In later sentence regions, however, a disadvantage for OS sentences emerged regardless of context or object type, though less pronounced in Experiment 3. Thus, discourse properties licensing OS order temporarily alleviated, but did not eliminate, processing difficulty for OS sentences.

Our finding of facilitative context effects in early sentence regions is consistent with prior studies showing that context can exert effects while a noncanonical sentence is being processed

(Kaiser & Trueswell, 2004) and in particular before the main verb is reached (Bornkessel et al., 2003; Koizumi & Imamura, 2017; Yano & Koizumi, 2018). Our results also confirm studies which demonstrated that effects of context on online processing are not restricted to cases of syntactic ambiguity, but affect unambiguous sentences as well (Grodner et al., 2005).

Results across studies differ with respect to whether the facilitative effect of context is temporary, with a disadvantage for noncanonical sentences emerging later in the sentence regardless of whether context supports OS order or not (as in Koizumi & Imamura, 2017, and the experiments reported here), or whether the facilitative effect of context persists until the end of the sentence (Kaiser & Trueswell, 2004). Note, however, that Kaiser and Trueswell (2004) examined Finnish, an SVO language, whereas Koizumi and Imamura (2017) and the experiments reported here tested SOV languages and used structures in which object and subject NP were processed before reaching the main verb.

In contrast to the online data, discourse properties affected comprehension accuracy only in a very limited way. Experiment 1 showed no effect of context on comprehension accuracy at all. In Experiment 2, accuracy was slightly higher for OS sentences with a demonstrative object than for OS sentences with a definite object, but only when the question asked for the agent of the sentence. In this case, accuracy reached a value of 84%, which is not much below the 90% reached by SO sentences followed by an agent question. Using the same contexts and demonstrative objects, but syntactically less complex questions, Experiment 3 revealed high accuracy values for both agent and patient questions, with only a small disadvantage for OS sentences in comparison to SO sentences. We therefore conclude that the low accuracy for comprehension questions observed in some conditions in Experiments 1 and 2 cannot be attributed to systematic misinterpretation, in line with conclusions drawn in Paolazzi et al. (2019), Bader and Meng (2018), and Cutter et al. (2022). We will advance a proposal concerning the source of the observed misinterpretation errors in the next section.

Regarding the online data, an account is needed why OS sentences induce increased processing cost even when OS order was as acceptable as SO order, as confirmed by the pretest results of Experiment 2. Current discussions of sentence complexity have revolved around the distinction between memory-based and expectation-based explanations of comprehension cost (see the overview in Levy, 2013). As discussed in Levy (2013), sentence comprehension is subject to expectation and memory effects, sometimes even when considering a single construction type (for German, see Levy & Keller, 2013, and Vasishth & Drenhaus, 2011). The question thus is what type of effect(s) we are seeing in our reading time data.

According to surprisal theory (Levy, 2008) – an influential expectation-based model – the processing cost associated with a word is higher the less expected the word is given the preceding context. As research on English subject and object relative clauses has shown, surprisal theory typically predicts a reading time penalty at the beginning of sentences with

noncanonical word order whereas at later sentence positions either no difference or even an advantage for noncanonical sentences is predicted. Applied to German SO or OS main clauses, surprisal theory makes three predictions that are similar to the predictions for English relative clauses. First, sentences start much more often with a subject than an object. In her corpus study of written German, Hoberg (1981, p. 162) found that about 58% of all main clauses started with the subject, but only about 3.5% started with an object (the rest started with an adverbial or a predicate nominal). Reading times should therefore be higher for OS than for SO sentences on the first NP unless the preceding context makes the OS structure as likely as the SO structure. Second, the likelihood of seeing a subject is high after having seen an object, because almost all sentences with an object also have a subject. The likelihood of seeing an object after having seen a subject is not as high because subjects do not have to be accompanied by an object, as in intransitive sentences, for example. Reading times should therefore be higher for SO than for OS sentences on the second NP. Third, after both subject and object have been encountered, no further differences between SO and OS sentences are expected because the likelihood of what follows subject and object should be independent of the order between them.

The reading time data yielded by the three experiments reported in this paper are only in partial agreement with these three predictions. In accordance with surprisal theory, we found a smaller or no reading time penalty at the beginning of OS sentences when the context and the object's determiner supported the use of OS sentences. On the second NP, there was either no difference or a penalty for OS sentences. For the rest of the sentences, there was also either no difference or a penalty for OS order, and in all three experiments, the disadvantage for OS sentences was particularly pronounced on the clause-final verb. In sum, while surprisal theory is compatible with the early effects of context that we found, the increased reading times for OS sentences at later positions cannot be explained in terms of surprisal.

Note that our results differ from the findings of Grillo et al. (2018) on the relationship between German active and passive clauses. In two selfpaced reading experiments, Grillo et al. found faster reading times for the second NP in passive clauses, that is, the NP within the by-phrase of a passive clause, than the second NP in active SO sentences, that is, the direct object. The authors explain this finding in terms of surprisal – the second NP is highly predictable in a passive clause after the parser has already processed the passive auxiliary and the preposition *von* ('by'), whereas the direct object is not predictable. There are several possible non-exclusive reasons for the lack of a comparable effect in our experiments – slower reading times on the subject in OS sentences than on the object in SO sentences. First, although the subject is highly expected in an OS sentence, its position is variable; it can follow the finite verb immediately but it can also be separated from it by adverbials. Furthermore, passive clauses are less marked than OS clauses because they still start with the subject.

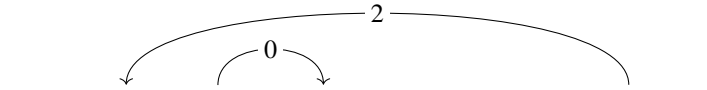
Memory-based explanations are the major alternative to expectation-based explanations. According to the Dependency Locality Theory (DLT) (Gibson, 2000), two memory components contribute to processing cost, storage cost and integration cost. Storage costs reflect open dependencies that are awaiting completion and must therefore simultaneously be stored in working memory. For SO and OS sentences as investigated in Experiments 1–3, storage costs differ at the initial NP and the finite verb depending on order. At these early positions, there is a storage cost of one for SO sentences because a main verb is still needed for sentence completion. OS sentences, in contrast, come with a storage cost of two because in addition to a main verb, a subject NP is needed in a complete sentence. When the second NP is processed, the subject is found. From the second NP onward, SO and OS sentences are therefore both associated with a storage cost of one until the end of the sentence is reached and the main verb leads to sentence completion.

Integration costs, the second processing cost component postulated by the DLT, measure the amount of processing needed to integrate each newly encountered input word into the structural representation built thus far. A main contributor to integration cost is the number of new discourse referents intervening between the first and the second element of a syntactic dependency. For the sentences under consideration, (8) shows the two dependencies relevant for the current discussion.[2]

The subject and the finite verb are connected by a dependency because the finite verb assigns nominative case to the subject and agrees with it in number and person. The dependency between object and main verb is necessary because the verb assigns case to its object.

(8) a.

b.



Since nothing intervenes between subject and finite verb – both are adjacent to each other in SO and OS sentences – no order-dependent processing differences are predicted at early sentence positions. The dependency between object and main verb has to be computed at the end of the sentence when the main verb is encountered in the input. In an SO sentence, only adverbial material intervenes between object and main verb so that the integration depends on how many

[2] Alternatively, one could make both subject and object depend on the main verb, which assigns a thematic role to both of them (see the dependency structure encoded in the Universal Dependency Project, https://universaldependencies.org/de/index.html). In this case, integration costs would not differ depending on the order of subject and object.

new discourse referents are introduced by the adverbial(s). In example (8), one new referent is introduced (*Maske* 'mask'). In OS sentences, the same adverbial material intervenes between object and main verb, but the subject also intervenes, so integration costs are always higher by one in OS sentences in comparison to SO sentences, independent of the complexity of the following adverbial material. Integration cost at the sentence-final main verb is thus higher in OS than in SO sentences.

In sum, a memory-based explanation along the lines of the DLT of Gibson (2000) could explain several of the findings yielded by the present experiments. Storage costs are higher for OS sentences at the beginning of a sentence and equal for SO and OS sentences at later positions. Integration costs, on the other hand, are equal up to the clause-final verb where they are higher for OS sentences. However, certain findings remain unaccounted for under the DLT. First, the discourse-based attenuation of the reading time disadvantage for OS sentences is not captured by the memory-based DLT. Here, one could combine Surprisal Theory and DLT to also account for this finding, as has been done for other structures by, e.g., Levy and Keller (2013); Vasishth and Drenhaus (2011). The one remaining finding that is neither accounted for by the DLT nor by Surprisal Theory is the finding of prolonged reading times for OS sentences from the second NP onward up to the clause-final verb in Experiments 1 and 2 but not in Experiment 3, where the OS disadvantage showed up only on the clause-final verb. This could indicate that the complexity of the secondary task required after reading a sentence may also contribute to specific differences in reading time patterns. Although all three experiments reported here used the same task of answering wh-questions, the questions themselves were less complex in Experiment 3 than in Experiments 1 and 2. This raises the possibility that small and localized differences between SO and OS sentences get magnified when participants become aware that sentences are followed by questions that are sometimes difficult to answer. More research examining different task requirements is needed to clarify this issue.

5.2 The source of misinterpretation errors

Our experiments revealed a robust effect of question type on the accuracy of comprehension questions. For both SO and OS sentences – though for OS sentences to a lesser extent – the number of misinterpretation errors varied depending on the particular form of the question. For ease of reference, the accuracy results are summarized in (9).

(9)

Target sentence:	SO		OS		OS		SO
Question:	Wh-Agent		Wh-Agent		Wh-Patient		Wh-Patient
Experiment 1	89	>	69	>	62	>	52
Experiment 2	91	>	79	>	65	>	53
Experiment 3	86	>	78	≈	81	<	85

In Experiments 1 and 2, accuracy was highest for SO sentences followed by an agent question and lowest for SO sentences followed by a patient question. Accuracy for OS sentences was in between, and again agent questions resulted in higher accuracy than patient questions. Experiment 3 used questions with a single argument only and found equally high accuracy for agent and patient questions. In a mixed-effects model with Question Type (agent vs. patient question), Order (SO vs. OS), and Question Complexity (two arguments, Experiments 1/2 vs. one argument, Experiment 3) as fixed effects and participants and sentences as random effects, all main effects and all interactions except the interaction of Order and Question Complexity were significant. Pairwise comparisons conducted to explore the crucial interaction between Question Type and Question Complexity revealed that for two-argument questions, agent questions resulted in significantly more correct answers than patient questions (81% vs. 58%; $\beta = 1.110$; SE $= 0.171$; $z = 6.49$, $p < 0.01$) whereas for one-argument questions, there was no significant difference between agent and patient questions (82% vs. 83%; $\beta = –0.060$; SE $= 0.202$; $z = –0.30$, n.s.).

The question now is what causes the pattern of misinterpretation errors. According to the GE model, misinterpretation errors result from applying heuristics that are fallible in case of noncanonical sentences. As discussed above, the relatively high accuracy values for OS sentences in some conditions are not easy to account for in the GE model. Furthermore, we did not find the relationship between correctness and reading times expected under the GE account. However, as pointed out by a reviewer, it cannot be ruled out that the predicted relationship between reading times and correctness is obscured by additional processes that intervene between reading the sentence and answering the comprehension question. For example, participants may read sentences with a fast pace but then pause and silently “re-read” the sentence before giving their answer, thereby preventing any simple relationship between reading times and question accuracy. We must leave it to future research whether an amendment of the GE model along these lines is viable.

Overall, our findings fit well with accounts that ascribe misinterpretation errors to fallible memory mechanisms – to degrading memory representations (Paolazzi et al., 2019, 2021) or to fallible retrieval operations (Bader & Meng, 2018; Meng & Bader, 2021). A detailed account of what these retrieval operations and memory representations are and why they often fail in the case of noncanonical sentences is still lacking, however. As a first step toward a comprehensive retrieval account, we will now specify the individual processing steps that lead from a wh-question to the retrieval of a phrase from the memory representation of a target sentence. We first note that accessing the answer directly by means of a simple cue-based memory retrieval would not account for the error pattern. Accessing the answer using semantic role as a retrieval cue is not compatible with finding many errors with patient questions in Experiments 1 and 2 but not in Experiment 3. Syntactic function cannot act as a simple retrieval cue either, because a wh-subject matches the target sentence’s subject when the question is in the active voice but its object when the question is in the passive voice. A retrieval cue “subject” would thus be underspecified.

In the case of two-argument questions, question and target sentence could be matched sequentially – first NP with first NP and second NP with second NP. A corresponding parallelism effect is clearly visible for SO sentences: accuracy was very high for agent questions (= SO order) and very low for patient questions (= OS order). Interestingly, however, OS sentences did not show a parallelism effect. For them, accuracy was low in general, and it was even somewhat lower for the syntactically parallel OS patient questions.

Thus, a more intricate way of matching questions to target sentences is needed. An important clue as to how such a matching algorithm could work is provided by the finding that accuracy was always best when the wh-phrase was a subject: (i) In Experiments 1 and 2, accuracy for SO sentences was much higher for agent questions (wh-phrase = subject) than for patient questions (wh-phrase = object). For OS sentences, accuracy was also higher when the question asked for the subject, although the difference was smaller. (ii) When the patient was queried using a subject wh-phrase, as in the passive questions of Experiment 3, accuracy was equally high as for agent questions. (iii) Throughout, reaction times were substantially higher for OS patient questions than for SO agent questions.

In order to capture the better performance with subject questions, we propose that the subject of the question – whether it is the wh-phrase or not – serves as starting point when matching the question to the target sentence. After the subject has been matched, further arguments are matched if necessary. The overall algorithm for retrieving the answer given a wh-question is given in (10).

(10) a. Step 1: Align the subject NP of the question with the corresponding phrase of the target sentence. If the subject of the question is a wh-phrase, go to Step 3, otherwise go to Step 2.
b. Step 2: Align the object NP of the question with the object phrase of the target sentence.
c. Step 3: Retrieve the phrase of the target sentence corresponding to the wh-phrase in the question.

A further finding, which is not yet accounted for by the question answering algorithm in (10), is that for both agent and patient questions with two arguments, accuracy was higher when the phrase of the target sentence corresponding to the wh-phrase appeared sentence-initially (SO with agent questions, OS with patient question). In order to capture this additional finding, we assume that accessing a phrase within working memory is easiest for phrases in sentence-initial position. Independent evidence for this assumption comes from research on anaphora resolution (e.g., Gernsbacher & Hargreaves, 1988; Gordon et al., 1993) and research on sentence interpretation, including Ferreira (2003, p. 175), who found that participants were more accurate naming the agent than the patient for active sentences (agent-initial), and more accurate naming the patient than the agent for passives (patient-initial).

To see how the proposed algorithm works, let us work through the different combinations of target sentences and question types, as illustrated in **Table 13**. Agent questions following an SO target sentence are the simplest case. Because the subject of the question and the subject of the target sentence both occur sentence-initially, locating the phrase corresponding to the question's subject is straightforward. Since in agent questions the subject is the wh-phrase, Step 2 can be skipped. The final step – retrieving the correct answer – is easy too, because the phrase in the target sentence that corresponds to the wh-phrase in the question occurs in sentence-initial position.

Table 13: Illustration of the wh-question answering algorithm in (10).

		SO target sentence	*OS target sentence*
Agent question	Target:	subject object	object subject
	Question:	**wh-subject** object	**wh-subject** object
Patient question	Target:	subject object	object subject
	Question:	wh-object **subject**	wh-object **subject**

Consider next OS sentences followed by an agent wh-question. The first step is more involved in this case because the subject now occurs sentence-finally in the target sentence. Step 2 can be skipped again. In order to get the correct answer at Step 3, the sentence-final subject must be retrieved from the target sentence. Given the evidence discussed above that sentence-initial constituents are more accessible than sentence-final ones, retrieving the subject from an OS target sentence is a more complex operations than retrieving the subject from an SO target sentence.

We now turn to patient questions, which elicit more errors on comprehension questions and higher reaction times. The two cases to consider are shown in the lower part of **Table 13**. Here, the subject occurs in final position, so locating the subject within the question in order to start the matching process is more complex than in the case of agent questions. Furthermore, since the subject is not the wh-phrase, the alignment established during the first step does not already identify the phrase needed for the answer. The wh-phrase is processed during the second step, when the objects are aligned. Only then does the information needed for retrieving the answer become available. We hypothesize that the elevated reaction times for patient questions compared to agent questions reflect the need to first match the subject before the wh-object can be used to identify the answer phrase. As shown in **Table 13**, for patient questions OS sentences are at an advantage because the phrase that needs to be retrieved occurs sentence-initially, and also because of parallelism between question and target sentence.

In contrast to two-argument questions, one-argument questions were answered with the same high accuracy independently of whether the question asked for the agent or the patient. This follows from the proposed question answering algorithm because the wh-phrase in one-argument questions is the subject in both agent and patient questions.

Before concluding, we should point out that our account of question answering does not imply that object questions are always difficult to process. We consider it likely that the difficulties seen with object questions in Experiments 1 and 2 are also due to having the indefinite pronoun *jemand* ('someone') as subject. Because the indefinite pronoun is semantically impoverished, it is not of great help when matching the subject of the question to the subject of the target sentence, as required in Step 1 of the proposed algorithm. With a lexical subject instead, this step is much easier to accomplish, as illustrated in (11). Here, we would expect that both agent and patient questions are answered with high accuracy. Crucially, the patient question should still need more time to be answered than the agent question.

(11) a. Target sentence: Der König hat den Botschafter trotz der Maske erkannt.
'The king recognized the ambassador despite the mask.'
b. Agent-Question: Wer hat den Botschafter erkannt?
'Who recognized the ambassador?'
c. Patient-Question: Wen hat der König erkannt?
'Who did the king recognize?'

In sum, we hypothesize that providing an answer to a wh-question proceeds in several steps that are fallible to varying degrees, depending on properties of the question and properties of the target sentence. We have specified the individual steps in a qualitative way and must leave a formal implementation allowing for quantitative predictions as work for future research.

5.3 Conclusion

In three experiments, we have investigated the comprehension of sentences with either canonical SO or noncanonical OS word order in context, using selfpaced reading as online and comprehension questions as offline measures. The reading time results show an overall disadvantage for OS sentences that is attenuated when the context or the object's determiner favors OS order. The offline results revealed that the accuracy of answering a question depended strongly on the syntactic form of the question. To capture the pattern of misinterpretation errors, we have proposed that participants understood the sentences correctly most of the time, but sometimes retrieved the false answer due to difficulties with matching question and target sentence.

Data accessibility statement

Stimuli, datasets and analysis code are available at: https://osf.io/h9a5w/?view_only=9edcb2ea311d40c8a7e5786a9e823c36

Competing interests

The authors have no competing interests to declare.

Author contributions

Both authors designed and conducted the research, analysed the data and wrote the paper. The first author planned the statistical analysis and implemented the experiments on PsychoPy and IBEX farm.

References

Bader, M., & Meng, M. (1999). Subject-object ambiguities in German embedded clauses: An across-the-board comparison. *Journal of Psycholinguistic Research*, *28*(2), 121–143. DOI: https://doi.org/10.1023/A:1023206208142

Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1286–1311. DOI: https://doi.org/10.1037/xlm0000519

Bader, M., & Portele, Y. (2019). Givenness and the licensing of object-first order in German: The effect of referential form. In A. Gattnar, R. Hörnig, M. Störzer, & S. Featherston (Eds.), *Proceedings of Linguistic Evidence 2018. Experimental Data Drives Linguistic Theory* (pp. 208–228). University of Tübingen, online publication system. DOI: https://doi.org/10.15496/publikation-32622

Bader, M., & Portele, Y. (2021). Discourse and form constraints on licensing object-first sentences in German. *Languages*, *6*(2), 82. DOI: https://doi.org/10.3390/languages6020082

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. *arXiv.org preprint – arXiv:1506.04967 [stat.ME]*. Retrieved from https://arxiv.org/abs/1506.04967v1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Bornkessel, I., & Schlesewsky, M. (2006). The role of contrast in the local licensing of scrambling in German: Evidence from online comprehension. *Journal of Germanic Linguistics*, *18*(1), 1–43. DOI: https://doi.org/10.1017/S1470542706000018

Bornkessel, I., Schlesewsky, M., & Friederici, A. D. (2003). Contextual information modulates initial processes of syntactic integration: The role of inter-versus intrasentential predictions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 871. DOI: https://doi.org/10.1037/0278-7393.29.5.871

Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, *86*(1), 168–213. DOI: https://doi.org/10.1353/lan.0.0189

Burmester, J., Spalek, K., & Wartenburger, I. (2014). Context updating during sentence comprehension: The effect of aboutness topic. *Brain and Language*, *137*, 62–76. DOI: https://doi.org/10.1016/j.bandl.2014.08.001

Christensen, R. H. B. (2019). *ordinal—Regression models for ordinal data.* (R package version 2019.4-25. http://www.cran.r-project.org/package = ordinal/)

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 817–828. DOI: https://doi.org/10.1080/17470218.2015.1134603

Christianson, K., & Luke, S. G. (2011). Context strengthens initial misinterpretations of text. *Scientific Studies of Reading*, *15*(2), 136–166. DOI: https://doi.org/10.1080/10888431003636787

Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 538–544. DOI: https://doi.org/10.1037/a0018027

Chromý, J. (2021). When readers fail to form a coherent representation of garden-path sentences. *Quarterly Journal of Experimental Psychology*, *75*(1), 169–190. DOI: https://doi.org/10.1177/17470218211037152

Cutter, M. G., Paterson, K. B., & Filik, R. (2022). Online representations of non-canonical sentences are more than good-enough. *Quarterly Journal of Experimental Psychology*, *75*(1), 30–42. DOI: https://doi.org/10.1177/17470218211032043

Dempsey, J., & Brehm, L. (2020). Can propositional biases modulate syntactic repair processes? Insights from preceding comprehension questions. *Journal of Cognitive Psychology*, 1–10. DOI: https://doi.org/10.1080/20445911.2020.1803884

Drummond, A., Von Der Malsburg, T., Erlewine, M. Y., Yoshida, F., & Vafaie, M. (2016). *Ibex Farm.* Retrieved from https://github.com/addrummond/ibex

Fanselow, G., Lenertová, D., & Weskott, T. (2008). Studies on the acceptability of object movement to Spec, CP. In A. Steube (Ed.), *The discourse potential of underspecified structures* (Vol. 8, pp. 413–438). De Gruyter. DOI: https://doi.org/10.1515/9783110209303.4.413

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203. DOI: https://doi.org/10.1016/S0010-0285(03)00005-7

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. DOI: https://doi.org/10.1111/j.1749-818X.2007.00007.x

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

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*(6), 699–717. DOI: https://doi.org/10.1016/0749-596X(88)90016-2

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain. Papers from the first Mind Articulation Project Symposium* (pp. 95–126). Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/3654.003.0008

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. DOI: https://doi.org/10.1073/pnas.1216438110

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*(3), 311–347. DOI: https://doi.org/10.1207/s15516709cog1703_1

Grillo, N., Alexiadou, A., Gehrke, B., Hirsch, N., Paolazzi, C., & Santi, A. (2018). Processing unambiguous verbal passives in German. *Journal of Linguistics*, *55*(3), 1–40. DOI: https://doi.org/10.1017/S0022226718000300

Grodner, D., Gibson, E., & Watson, D. (2005). The influence of contextual contrast on syntactic processing: Evidence for strong-interaction in sentence comprehension. *Cognition*, *95*(3), 275–296. DOI: https://doi.org/10.1016/j.cognition.2004.01.007

Hoberg, U. (1981). *Die Wortstellung in der geschriebenen deutschen Gegenwartssprache*. München: Hueber.

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes*, *26*(3), 376–405. DOI: https://doi.org/10.1080/01690965.2010.492642

Hopp, H., Bail, J., & Jackson, C. N. (2020). Frequency at the syntax–discourse interface: A bidirectional study on fronting options in L1/L2 German and L1/L2 English. *Second Language Research*, *36*(1), 65–96. DOI: https://doi.org/10.1177/0267658318802985

Imamura, S., Sato, Y., & Koizumi, M. (2016). The processing cost of scrambling and topicalization in Japanese. *Frontiers in Psychology*, *7*, 531. DOI: https://doi.org/10.3389/fpsyg.2016.00531

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. DOI: https://doi.org/10.1037/0033-295X.87.4.329

Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, *94*(2), 113–147. DOI: https://doi.org/10.1016/j.cognition.2004.01.002

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, *69*(3), 1013–1040. DOI: https://doi.org/10.1080/17470218.2015.1053951

Koizumi, M., & Imamura, S. (2017). Interaction between syntactic structure and information structure in the processing of a head-final language. *Journal of Psycholinguistic Research*, *46*(1), 247–260. DOI: https://doi.org/10.1007/s10936-016-9433-3

Kristensen, L. B., Engberg-Pedersen, E., & Poulsen, M. (2014). Context improves comprehension of fronted objects. *Journal of Psycholinguistic Research*, *43*(2), 125–140. DOI: https://doi.org/10.1007/s10936-013-9241-y

Lenerz, J. (1977). *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Narr.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. DOI: https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R. (2013). Memory and surprisal in human sentence comprehnsion. In R. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). New York: Psychology Press.

Levy, R., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language, 68*(2), 199–222. DOI: https://doi.org/10.1016/j.jml.2012.02.005

Meng, M., & Bader, M. (2021). Does comprehension (sometimes) go wrong for noncanonical sentences? *Quarterly Journal of Experimental Psychology*, *74*(1), 1–28. DOI: https://doi.org/10.1177/1747021820947940

Meng, M., Bader, M., & Bayer, J. (1999). Die Verarbeitung von Subjekt-Objekt-Ambiguitäten im Kontext. In I. Wachsmuth & B. Jung (Eds.), *KogWiss99. Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft Bielefeld, 28. September – 1. Oktober 1999* (p. 244–249). St. Augustin: Infix Verlag.

Paolazzi, C. L., Grillo, N., Alexiadou, A., & Santi, A. (2019). Passives are not hard to interpret but hard to remember: Evidence from online and offline studies. *Language, Cognition and Neuroscience*, *34*(8), 991–1015. DOI: https://doi.org/10.1080/23273798.2019.1602733

Paolazzi, C. L., Grillo, N., & Santi, A. (2021). The source of passive sentence difficulty: Task effects and predicate semantics, not argument order. In *Passives Cross-Linguistically* (pp. 359–393). Brill. DOI: https://doi.org/10.1163/9789004433427_012

Peirce, J. W. (2007). PsychoPy – psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. DOI: https://doi.org/10.1016/j.jneumeth.2006.11.017

Qian, Z., Garnsey, S., & Christianson, K. (2018). A comparison of online and offline measures of good-enough processing in garden-path sentences. *Language, Cognition and Neuroscience*, *33*(2), 227–254. DOI: https://doi.org/10.1080/23273798.2017.1379606

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rosenbach, A. (2005). Animacy versus weight as determinants of grammatical variation in English. *Language*, *81*(3), 613–644. DOI: https://doi.org/10.1353/lan.2005.0149

Schlesewsky, M., Fanselow, G., Kliegl, R., & Krems, J. (2000). The subject preference in the processing of locally ambiguous Wh-questions in German. In B. Hemforth & L. Konieczny (Eds.), *German sentence processing* (pp. 65–93). Dordrecht: Kluwer. DOI: https://doi.org/10.1007/978-94-015-9618-3_3

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension. The integration of habits and rules*. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/6184.001.0001

Vos, S. H., & Friederici, A. D. (2003). Intersentential syntactic context effects on comprehension: The role of working memory. *Cognitive Brain Research*, *16*(1), 111–122. DOI: https://doi.org/10.1016/S0926-6410(02)00226-4

Vasishth, S., & Drenhaus, H. (2011). Locality in German. *Dialogue & Discourse*, *2*(1), 59–82. DOI: https://doi.org/10.5087/dad.2011.104

Weskott, T., Hörnig, R., Fanselow, G., & Kliegl, R. (2011). Contextual licensing of marked OVS word order in German. *Linguistische Berichte*, *225*, 3–18.

Wonnacott, E., Joseph, H. S. S. L., Adelman, J. S., & Nation, K. (2016). Is children's reading "good enough"? Links between online processing and comprehension as children read syntactically ambiguous sentences. *Quarterly Journal of Experimental Psychology*, *69*(3), 855–879. DOI: https://doi.org/10.1080/17470218.2015.1011176

Yano, M., & Koizumi, M. (2018). Processing of non-canonical word orders in (in)felicitous contexts: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *33*(10), 1340–1354. DOI: https://doi.org/10.1080/23273798.2018.1489066