The type of garden-path matters: When readers fail to form a coherent representation of garden-path sentences

Author

Jan Chromý

Author's affiliation

Institute of the Czech Language and Theory of Communication

Faculty of Arts, Charles University, Prague, Czech Republic

Address: nám. Jana Palacha 2, Praha 1, 116 38, Czech Republic

Corresponding Author

Jan Chromý (jan.chromy@ff.cuni.cz)

Postal address: Jan Chromý, Faculty of Arts, nám. Jana Palacha 2, Praha 1, 116 38, Czech Republic

Online materials (stimuli, data, preregistrations)

https://osf.io/bjas8/

The type of garden-path matters: When readers fail to form a coherent representation of garden-path sentences

Abstract

Various studies within the Good-Enough Approach observe that people often make errors in answering comprehension questions after reading garden-path sentences such as *While Anna dressed the baby played in the crib* (e.g. Christianson, Hollingworth, Halliwell, & Ferreira, 2001). Recently (Slattery, Sturt, Christianson, Yoshida, & Ferreira, 2013), it has been claimed that readers form a full syntactic analysis of these sentences, but they do not completely prune the original misanalysis. This paper presents evidence that there are important differences between the various types of garden-path structures in terms of their final representations. The main finding of the Good-Enough Approach – that the comprehension questions targeting the initial misanalysis yield significantly higher rates of incorrect answers after garden-path sentences, in comparison to after control sentences – was replicated here in three self-paced reading experiments. However, these experiments show a similar pattern of results for other comprehension questions, such as questions targeting an analysis that is not syntactically licensed at any point of processing. The results of this paper, together with previous studies, point out that there is a range of difficulty levels for garden-path structures, which is related to a range of outcome representations. Some garden-paths are easy to process and the resulting representation is typically correct. Others are harder and result in a lingering misanalysis. In the hardest cases, readers may often end up with a disrupted and incoherent representation of the sentence.

Keywords

sentence processing; garden-path; ambiguity; Czech; self-paced reading

Word count: 14 911

1 Introduction

Garden-path (GP) sentences (e.g. *While Anna dressed the baby played in the crib*) have been one of the most widely studied linguistic structures since the beginning of modern psycholinguistic research (Bever, 1970; Frazier & Fodor, 1978). Researchers have been interested in various aspects of the processing of these sentences, such as whether or not readers activate only one analysis at a time (MacDonald, Pearlmutter, & Seidenberg, 1994), if and how the parsing of these sentences is influenced by non-syntactic factors such as semantics (Sturt, 2007), plausibility (e.g. Pickering & Traxler, 1998), subcategorization frequency (e.g. Pickering, Traxler, & Crocker, 2000), and context (Altmann, Garnham, & Dennis, 1992). For a long time, research on garden-path processing was focusing on how a complete analysis is formed and what factors influence this process. Cases such as misunderstandings or incomplete processing were typically overlooked (with exceptions such as Fodor & Inoue, 1994, 1998).

However, Christianson, Hollingworth, Halliwell, & Ferreira (2001) showed that there seems to be a systematic tendency to analyze GP sentences such as *While the man hunted the deer that was brown and graceful ran into the woods* incorrectly. This finding – among others – led Ferreira, Christianson and their colleagues to the idea of Good-Enough Processing (e.g. Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007; Karimi & Ferreira, 2016; Christianson, 2016). Researchers working within this approach claim that language comprehension (and sentence processing) is, at least sometimes, only partial and that semantic representations are often incomplete. More precisely, they claim that apart from algorithmic processes that are responsible for computing syntactic structures, people often use simple heuristics which are fast and frugal. According to this view, such heuristics help us to save processing resources – they are good-enough in the sense that they give us an approximate message and thus typically lead to communicative success. One of the well-studied structures in the Good-Enough Approach have been the GP sentences; it has been shown that the initial misanalysis of these sentences tends to linger.

1.1 Good-Enough Approach and the processing of garden-path sentences

In the seminal paper in this line of research, Christianson et al. (2001) conducted a series of five experiments. Participants read sentences like (1a) and (1b).

(1a) While the man hunted the deer that was brown and graceful ran into the woods.

(1b) The deer that was brown and graceful ran into the woods while the man hunted.

Sentences like (1a) are locally ambiguous – they contain a GP structure. Sentences like (1b) serve as control sentences which have the same meaning as (1a), but they are not locally ambiguous. The local ambiguity of (1a) lies in the region *the deer,* which is understood as the object of the verb *hunted* in the first analysis. This analysis lasts until the participant encounters the verb *ran*. After that, the first analysis should be dismissed, and the sentence reanalyzed so that the region *the deer* is understood as the subject of the verb *ran*. The authors were interested mainly in whether participants really dismiss the initial (GP) interpretation, or if the initial interpretation ('Man hunted the deer') lingers, i.e. is continuously active. After reading each experimental or control sentence, the participants had to answer a yes-no comprehension question targeting the initial misanalysis (e.g. *Did the man hunt the deer?*) and state how certain they are about their choice. The study found that there is a general tendency to (incorrectly) respond "yes" after GP sentences: the comprehension question after the sentence (1a) yielded 75% incorrect responses whereas the same question after the control non-garden-path (non-GP) sentence (1b) yielded 49% incorrect responses. Similar results were also found for sentences like (2a) and (2b) in which there was a reflexive absolute transitive verb (such as *dressed*):

(2a) While Anna dressed the baby that was small and cute spit up on the bed.

(2b) The baby that was small and cute spit up on the bed while Anna dressed.

A comprehension question *Did Anna dress the baby?* yielded 65.6% incorrect answers after sentence (2a) and only 12.5% incorrect answers after sentence (2b). The authors thus concluded that the initial (GP) interpretation was often active even after reading the whole sentence, and that the resulting representation thus did not match the real sentence content.

In a follow-up study, Christianson, Williams, Zacks, & Ferreira (2006) analyzed differences between younger and older speakers. They found similar general effects as in the first study. Interestingly, they also found that the older group of speakers showed an even higher rate of incorrect answers to questions following GP sentences than the younger group. The authors attribute this effect to older adults' increased reliance on heuristic-like good-enough processing. Similar effects were later also found by Patson, Darowski, Moon, & Ferreira (2009), who used a paraphrasing task instead of yes-no questions.

In later studies focusing on the Good-Enough Approach, it has been reported that higher intelligence and higher processing speed are related to a lower rate of incorrect answers (Engelhardt, Nigg, & Ferreira, 2017) and that there is no connection between rereading measures and comprehension accuracy (Christianson, Luke, Hussey, & Wochna, 2017). It is worth mentioning that the latter study analyzed a type of GP structure different from that of the previous studies (namely a reduced relative GP sentence such as *The player tossed the ball had interfered with the other team*) and they found a very low rate of correct answers (about 25%).

Another type of a GP structure – a noun phrase/sentence coordination ambiguity such as *The publisher called up the editor and the author refused to change the book's ending* – was analyzed in three experiments in Christianson & Luke (2011). The authors examined the role of preceding context, which was either neutral (such as *There was a public outcry against the publisher of a racy new novel*), GP-biased (e.g., *There was a public outcry against the author of a racy new novel*), or non-GP-biased (e.g., *There was a public outcry against the editor of a racy new novel*). The authors found that both reading times and response accuracy were influenced by preceding context and that the comprehension question wording may bias readers toward the incorrect interpretation even for the non-GP sentences. Interestingly, however, there was no difference in the response accuracy between the GP and non-GP sentences for neutral contexts (both GP and non-GP sentences yielded only 7% incorrect answers in the first experiment; and the ratio of incorrect answers between GP and non-GP sentences was 11% vs. 9% in the second experiment and 10% vs. 7% in the third one, with neither difference reaching significance). This result highlights that the response accuracy for different types of GP sentences may vary largely, probably due to the difference in the difficulty of repair of the initial, incorrect structure (cf. Fodor & Inoue, 1994, 1998).

An important study in the line of research on the lingering initial misanalysis is the paper by Slattery, Sturt, Christianson, Yoshida, & Ferreira (2013). The authors identified two possible explanations of the misinterpretation effects documented in the previous studies, namely (i) the syntactic representation is "incomplete, disconnected, or just plain wrong" (p. 105); and (ii) the parser initially creates an incorrect parse for the ambiguous material and "during reanalysis builds a new structure that is complete, fully specified, and faithful to the input, but that does not completely prune the original mis-analysis" (p. 106). They tested these possibilities in two eye-tracking experiments. In the first experiment, they examined participants' eye-movements while reading sentences such as:

(3a) After the bank manager telephoned David's father grew worried and gave himself approximately five days to reply. (Garden Path/Match)

(3b) After the bank manager telephoned David's mother grew worried and gave himself approximately five days to reply. (Garden Path/Mismatch)

(3c) After the bank manager telephoned, David's father grew worried and gave himself approximately five days to reply. (Non-Garden Path/Match)

(3d) After the bank manager telephoned, David's mother grew worried and gave himself approximately five days to reply. (Non-Garden Path/Mismatch)

For the reflexive pronoun (himself/herself), the authors observed an effect of gender mismatch for first-pass time, go-past time, and total time. However, the interaction between ambiguity and gender mismatch was not significant in any of these measures. The authors then expanded the analyzed region to the word following the reflexive pronoun (i.e. *approximately* in the examples above), but this did not yield a significant effect of the interaction between ambiguity and gender mismatch (except of the go-past time, which was however significant solely by items using an ANOVA). They interpret the fact that the processing of the pronoun in mismatch sentences like (3b) and (3d) takes longer than in match sentences as a sign that "the parser constructs a detailed syntactic structure" (p. 110).

In the second experiment, Slattery et al. (2013) tested whether the lingering misanalysis of a sentence might influence the processing of a sentence read immediately after. See example sentences 4a-d.

(4a) While Frank dried off the truck that was dark green was peed on by a stray dog. Frank quickly finished drying himself off then yelled out the window at the dog. (GP, plausible)

(4b) While Frank dried off, the truck that was dark green was peed on by a stray dog. Frank quickly finished drying himself off then yelled out the window at the dog. (non-GP, plausible)

(4c) While Frank dried off the grass that was dark green was peed on by a stray dog. Frank quickly finished drying himself off then yelled out the window at the dog. (GP, implausible)

(4d) While Frank dried off, the grass that was dark green was peed on by a stray dog. Frank quickly finished drying himself off then yelled out the window at the dog. (non-GP, implausible)

Among other effects on various regions, the authors found a significant interaction between sentence ambiguity and plausibility for first-pass times for the reflexive pronoun region (*himself off*). In other words, it took the participants significantly longer to initially process the pronoun and the following word in sentences (4a) in comparison to (4b) but there was no difference between (4c) and (4d). The authors interpret their findings under the lexically guided tree-adjoining grammar approach (Ferreira, Lau, & Bailey, 2004). They propose that a detailed hierarchical structure for GP sentences is formed, but in the resulting analysis, the ambiguous region both stays in the initial, incorrectly parsed position (where it competes with the correct structure) and is put in the correct, syntactically licensed position.

The findings of the Good-Enough Approach on GP sentences were further corroborated by Malyutina & den Ouden (2016), who used audio recorded stimuli and a sentence-picture matching task. In this task, participants heard various sentences and after hearing each sentence, they had to choose one picture out of three which best corresponded to the content of the sentence. They found that various linguistic factors play a role in the formation of the resulting interpretation of the sentence, for example sentence structure, verb type and semantic plausibility. Similarly to Christianson, Williams, Zacks, & Ferreira (2006), the authors found that older adults tended to answer more incorrectly than younger adults. Moreover, they argued that older speakers tend to maintain the initial representation without incorporating new information, whereas younger speakers tend to blend the two representations into one, even if this is not licensed by syntax. In other words, they claim that older speakers tend to represent the sentence *While Anna dressed the baby spit up on the bed* as a woman dressing a baby, but not the baby spitting on the bed, whereas younger speakers represented the sentence as a woman dressing a baby while it is spitting up on the bed.

Qian, Garnsey, & Christianson (2018) ran three experiments – two using self-paced reading and one using ERPs. Their aim was to test whether there is a relation between the time spent processing the disambiguating region in GP sentences and response accuracy. They predict that if the incorrect answers were caused by an incomplete reanalysis, the time spent processing the disambiguating region should be lower when the answer is incorrect. The authors did not observe such an effect. In the first self-paced reading experiment, there was no difference in reaction times (RTs on the disambiguating region, and in the second experiment, they found even an opposite effect, where the RTs were longer when the subsequent comprehension question was answered incorrectly. An analogous finding was made in the ERP experiment. As would be generally expected, the authors observed larger P600 for the disambiguating verbs in GP than in control sentences. However, the authors found no relationship between the P600 amplitude and the rate of incorrect answers to comprehension questions. In sum, Qian et al. (2018) interpret the findings as a sign of reanalysis being done even in the sentences which subsequently yielded incorrect answers on comprehension questions. They used these findings to argue against the idea of an incomplete reanalysis of the GP sentences.

Altogether, the results of the above-mentioned studies are convincing and highly convergent. It has been attested in numerous experiments using various methods that people tend to answer questions targeting the initial misanalysis incorrectly. However, two key limitations of these studies emerge, namely that (i) they only test a limited type of GP sentence, and (ii) the resulting representation is typically examined only through questions targeting the initial misanalysis. We will focus on these issues in the next section.

1.2 Limitations of previous studies

First, the types of GP sentences used in these studies are limited. Typically, sentences containing optionally transitive verbs (such as *to hunt* in the sentence *While the man hunted the deer ran into the woods*) or reflexive absolute transitive verbs (such as *to dress* in the sentence *While Anna dressed the baby played in the crib*) have been used. Additionally, the control sentences were created either by using a comma (*While Anna dressed, the baby played in the crib*) or by a different clause order (*The baby played in the crib while Anna dressed*). Only two studies used different structures, Christianson, Luke, Hussey, & Wochna (2017) examined the reduced relative (such as *The player tossed the ball had interfered with the other team*) and Christianson & Luke (2011) analyzed the noun phrase/sentence coordination ambiguity (such as *There was a public outcry against the publisher of a racy new novel*). The importance of examining a wider range of structures is motivated by the fact that the response accuracies seem to differ considerably both for the different GP structures and for the corresponding non-GP controls. For example, Christianson et al. (2001) found that the GP sentences containing the reflexive absolute transitive verbs yielded 57.3% incorrect answers compared to 11.5% for the controls. The reduced relative sentences examined in Christianson, Luke, Hussey, & Wochna (2017) yielded approximately 76% incorrect answers, whereas the non-GP controls yielded around 40–45% incorrect answers. And the response inaccuracy for the noun phrase/sentence coordination ambiguity (Christianson & Luke, 2011) was only around 7–10% for both GP and non-GP structures in cases where the preceding context was neutral. This variation may be attributed to the difference in the ease of recovery from the initial misanalysis for different GP structures (Fodor & Inoue, 1994, 1998; Van Dyke & Lewis, 2003). The resulting representation of different GP sentences may be quite different. It seems that the initial misanalysis may not linger at all (as the findings of Christianson & Luke (2011) would suggest for neutral contexts), in other cases, the initial misanalysis may linger, but the rest of the sentence may be represented faithfully to the input. It is also a possibility that there are GP sentences that are particularly difficult to process, to such an extent that some readers not only end up with a lingering misanalysis, but they fail to process the sentence fully and derive its correct representation.

Second, almost all the above-mentioned studies employ comprehension questions that target only one aspect of the understanding of the GP sentences, namely the initial misanalysis. There are only three exceptions. Christianson et al. (2001) used a question targeting the matrix clause in one of their five experiments, e.g. *Did the steak fall to the floor?* for sentences like *As Harry chewed the steak that was brown and juicy fell to the floor*. Interestingly, this question yielded a significantly higher rate of incorrect answers after the GP sentences than after the control sentences with a switched clause order (15% vs. 7.5% incorrect). Elsewhere, two other methods of testing comprehension have been employed: a picture matching task (Malyutina & den Ouden, 2016) and a paraphrasing task (Patson, Darowski, Moon, & Ferreira, 2009). However, the picture matching task offered three very explicit possible interpretations of the sentence and may thus be considered to exhibit the same flaws as the standard yes-no question methodology. In the paraphrasing task, participants were told to paraphrase the meaning of the sentence they just read (they were asked not to simply repeat the sentence and they were shown examples of unacceptable paraphrases – see Patson et al., 2009, p. 282). This is potentially more informative than yes-no questions, but it is still not very clear whether and to what extent it tests sentence comprehension, mere repetition, or reconstruction of the sentence content based on various memory cues.

The natural focus of the research on GP processing under the Good-Enough Approach has thus been on the initial misanalysis. However, one may question, what the resulting representation of the whole sentence is. Previous studies (e.g. Qian et al., 2018; Slattery et al., 2013) suggested that the reanalysis of the misparsed region is typically complete and full. For example, Slattery et al. (2013) claim that the parser creates a full syntactic structure, which also contains the initial misanalysis. In other words, a resulting interpretation of the sentence *While Harry dried off the truck that was dark green was peed on by a stray dog* would be simultaneously (i) *Harry dried off the truck*, (ii) *Harry dried off himself*, (iii) *the truck was peed on*, (iv) *it was a stray dog who peed on the truck*, or (v) *the truck was dark green*. However, the above-mentioned studies point to the fact that the GP processing may present a cognitively demanding task (depending on the type of GP structure). It is therefore an open question whether and how the cognitive effort needed for performing the reanalysis can affect the processing of the rest of the sentence. We may predict for GP structures where processing the meaning is particularly difficult, readers may be left with insufficient resources for the construction of an accurate representation of the other parts of the sentence. Even if the readers perform a complete and full reanalysis of the misparsed region, they may end up with an otherwise disrupted and confused representation of the sentence.

1.3 The present study

The present study aims to address the two limitations of the previous research on the good-enough processing of GP sentences. In three experiments in Czech, a different type of GP sentence is examined compared with the previous studies. To assess the comprehension of the GP sentences, various comprehension questions targeting different aspects of the sentence content are utilized, providing a means to test the resulting representations of the GP sentences.

The experiments use word-by-word self-paced reading and comprehension questions to examine processing of GP sentences, such as 5a, in comparison to non-GP control sentences, such as 5b.

(5a) garden-path condition (GP)

Kluci	honili	psa	a	kočk-u	v	podkroví
Boy-NOM.M.PL	chase-3PL.M.PST	dog-ACC.M.SG	and	cat-ACC.F.SG	in	attic-LOC.N.SG

znepokojovali	šediví	hlodavci.
worry-3PL.M.PST	grey-NOM.M.PL	rodents-NOM.M.PL

‘Boys chased a dog and grey rodents in the attic worried a cat.’

(5b) non-garden-path condition (non-GP)

Kluci	honili	psa	a	kočk-a	v	podkroví
Boy-NOM.M.PL	chase-3PL.M.PST	dog-ACC.M.SG	and	cat- NOM.F.SG	in	attic-LOC.N.SG

znepokojovala	šedivé	hlodavce.
worry-3SG.F.PST	grey-ACC.M.PL	rodents-ACC.M.PL

‘Boys chased a dog and a cat in the attic worried grey rodents.’

The GP structure in 5a is based on the coordination ambiguity where an NP following a conjunction (*a* in Czech meaning ‘and’) initially appears to be the second conjunct of a conjoined object in the first clause, but in fact is the object of the verb in the second clause. In other words, the parser should initially parse the NP *kočku* (‘a cat’ in the accusative case) as the object of the verb of the first clause (i.e. it forms the initial misanalysis meaning ‘Boys chased a dog and a cat’). This ultimately incorrect analysis lasts until the parser encounters the verb of the second clause (e.g. *znepokojovali*, ‘worried’) which should force the parser to conduct a reanalysis. In case the parsing proceeds correctly, it should come up with the final analysis where *kočku* is the object of the second clause (which has an OVS word order). In contrast, the form *kočka* ('a cat' in the nominative case) was used in the control sentences. This should effectively prohibit the parser to relate this NP to the first clause verb since the nominative case cannot stand as an object in Czech. Thus, the parser should instantly assume that *kočka* is the subject (and the first word) of the second clause.

One of the advantages of using sentences such as 5a and 5b is that the GP and control sentences are well-aligned and the RTs for individual regions are therefore directly comparable. Thus, it is also possible to relate the findings about the response accuracy on comprehension questions with the information on the on-line processing of the sentence. It should also be noted that the rules of Czech punctuation (Pravdová & Svobodová, 2014) explicitly forbid the use of a comma for separating the first and second clause in sentences like 5a and 5b. In other words, the GP effect should not be due to a non-presence of an expected comma.

A crucial difference between the GP sentence such as 5a and the noun phrase/sentence coordination ambiguity (such as *The publisher called up the editor and the author refused to change the book's ending*) analyzed by Christianson & Luke (2011) lies in the fact that the second clause of the Czech sentence has OVS word order. This word order can be used in Czech, but it is a marked word order typically used to focus the subject (Jasinskaja & Šimík, accepted). Siewierska & Uhlířová (1998) state (based on a corpus of approximately 30,000 clauses of written Czech) that OVS word order is used only in 14.6% of cases (compared to 63.1% for the canonical SVO word order). Therefore, the GP repair difficulty of 5a may be particularly strong because of the need to process a possibly unexpected OVS second clause. If an OSV clause was unexpected, the disambiguating verb of the second clause (e.g. znepokojovali 'worried' in 5a) would cause the parser to look primarily for a subject (which precedes the verb in the canonical and more expected SVO word order). However, the ambiguous NP (e.g. *kočku*) is in the accusative case and thus cannot be a valid subject of the second clause. In fact, there is no NP which would be a valid subject at that point of processing. On the other hand, the ambiguous NP would be a perfectly valid object of the first clause. Therefore, the parser may not have a tendency to link the second clause verb with the ambiguous NP as its object while processing the second clause verb and may proceed with the "attach anyway" strategy (Fodor & Inoue, 1998). Importantly, the following two regions (an adjective and a noun in nominative case that present a subject NP) may cause additional difficulties with processing if the parser previously gave up their search for the subject. In that case, the parser should anticipate an object to follow the verb (due to an expected SVO word order). However, the adjective and noun are both in nominative case and thus cannot constitute an object of this clause. In sum, 5a presents a GP structure which should be rather hard to repair.

Experiment 1 examines processing of GP sentences such as 5a and uses two types of comprehension questions, one targeting the initial misanalysis, the other targeting the analysis of the second clause. Experiment 2 uses a similar design to experiment 1, but uses two additional comprehension questions, one targeting the analysis of the first clause and the other targeting an analysis that should never occur during the reading of the sentence because it is not syntactically licensed at any point during processing. Experiment 3 is similar to Experiment 2, but uses an additional sentence condition where the initial misanalysis is semantically implausible.

In summary, the three experiments examine processing of a rather difficult GP structure together with response accuracy to various comprehension questions aimed at targeting not only the initial misanalysis, but also other aspects of the sentence representation. Thus, the experiments aim to investigate the idea that the resulting representation of the sentence may be disrupted due to a cognitive overload, which arises while conducting a reanalysis of the GP structure.

2 Experiment 1

The aim of Experiment 1 was to test the assumptions of the Good-Enough Approach to language comprehension on the processing of GP sentences in Czech. Similar to previous studies, the experiment analyzed responses to control questions presented after reading the GP and analogical non-GP sentences. Experiment 1 used word-by-word self-paced reading to measure RTs for each word in the sentence.

The stimuli lists and data used in the analysis are freely available on the Open Science Framework as supplementary materials at this link: https://osf.io/bjas8/

2.1 Method

2.1.1 Participants

Eighty-seven Charles University undergraduate students (72 female and 15 male; mean age 21.6 years) participated in Experiment 1. All participants were native speakers of Czech and participated for course credit.

2.1.2 Materials

Twenty-four experimental items were used in Experiment 1 (see Supplementary Materials for the whole list with English translation). Each item consisted of four conditions (2x2 factorial design) with two independent variables manipulated - sentence type and comprehension question type. Two sentence types (GP sentence 5a vs. non-GP sentence 5b) were used in each item, and each sentence type was followed by two types of yes-no question (one targeting the initial garden-path analysis as in 6a, and another targeting the resulting correct analysis of the sentence 6b). Each item thus comprised four conditions (5a+6a, 5a+6b, 5b+6a and 5b+6b). See Table 1 for an example item together with correct answers for each condition.

Table 1

Item example from Experiment 1 together with correct answer for each condition.

Condition	Sentence (Czech [English])		Comprehension question (Czech [English])		Correct answer
GP	5a	Kluci honili psa a kočku v podkroví znepokojovali šediví hlodavci. [Boys chased a dog and grey rodents in the attic worried a cat.]	6a	Honili kluci kočku? [Did the boys chase the cat?]	No
GP	5a	Kluci honili psa a kočku v podkroví znepokojovali šediví hlodavci. [Boys chased a dog and grey rodents in the attic worried a cat.]	6b	Znepokojovali hlodavci kočku? [Did the rodents worry the cat?]	Yes
non-GP	5b	Kluci honili psa a kočka v podkroví znepokojovala šedivé hlodavce. [Boys chased a dog and a cat in the attic worried grey rodents.]	6a	Honili kluci kočku? [Did the boys chase the cat?]	No
non-GP	5b	Kluci honili psa a kočka v podkroví znepokojovala šedivé hlodavce. [Boys chased a dog and a cat in the attic worried grey rodents.]	6b	Znepokojovali hlodavci kočku? [Did the rodents worry the cat?]	No

Each participant received only one condition of each item based on the Latin-square design and thus received six examples of each condition. There were also 120 filler sentences whose syntactic structure was different from the experimental sentences (none of which were GP sentences) and were also followed by a yes-no comprehension question. Forty-eight of these served as experimental items in another experiment. The comprehension questions were counterbalanced so that there was an even proportion of the negative and positive correct answers across the whole experiment.

2.1.3 Procedure

The experiment was conducted in the LABELS lab at Charles University. Participants were informed that the experiment consists of 144 sentences and that their task was to read word-by-word at their normal reading rate and that after each sentence a comprehension yes-no question appeared, which they had to answer by clicking the mouse. They were asked to use glasses in case they use them for reading. After this general introduction, they were seated in front of a computer, they completed a form containing several demographic questions. They started the experiment once this was completed. The experiment was programmed in IbexFarm 0.3.9 (http://spellout.net/ibexfarm/). At the beginning, participants read three practice sentences to get acquainted with the reading and answering procedure. The items and filler sentences were presented in randomized order. The experiment took about 20–25 minutes.

2.1.4 Data analysis

Before analyzing the results, the response accuracy on filler items was checked (experimental items were excluded from this analysis since their response accuracy was under investigation). The mean response accuracy was 93.62% (the median was 93.27%) and no participant had a response accuracy for filler items that was lower than 70%. No participant was thus excluded based on their response accuracy, and the high accuracy rates for filler items demonstrates that the participants read the sentences carefully.

The RTs were trimmed very conservatively; only those data points that were clearly discontinuous (less than 130 ms and more than 10 seconds) were excluded. This represented 0.13% of the data. Since the data were not normally distributed, the Box-Cox test (Box & Cox, 1964) was employed to establish the ideal data transformation method. This test yielded a score of $\lambda = -0.506$ which means that the ideal transformation would be inversely transformed squared RTs (1/sqrt(RTs)). The inversely transformed squared RTs were multiplied by -1000 so that the coefficients had the same sign and to avoid very small values or overly restricted ranges for the dependent variable values (a similar approach to inversely transformed RTs was adopted by Baayen & Milin, 2010).

Differences in RTs were analyzed for different sentence types. The analysis was run in R using linear-mixed effects models with the lme4 package (Bates, Mächler, Bolker, & Walker, 2014). The degrees of freedom and p-values were estimated using Satterthwaite's approximations from the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017). Three steps were followed in the analysis. First, a model was run predicting the inverse transformed squared RTs by length of word and item order (with a random intercept for participant). Second, residuals were extracted from that model. Third, the residuals were used as a dependent variable in a new model which included sentence type as a fixed effect and participant and item as random effects. Sentence type was sum coded (-0.5 for non-GP and +0.5 for GP sentences). The random effects structure (the inclusion of random slopes) was determined following Bates, Kliegl, Vasishth, & Baayen (2015). The beta estimates, standard errors, t-values, and p-values are reported (only for the statistically significant results). The regions for the sentence types are presented in Table 2. The target regions for the analysis were the disambiguating verb (region 8), the following adjective (region 9) and the sentence final noun (region 10). It was predicted that RTs on the region 8 in GP sentences should be elevated in comparison with this region in non-GP sentences. RTs on the next two regions were analyzed based on two reasons: (i) spillover effects may be expected there (see e.g. Christianson et al., 2017); (ii) the subject NP in regions 9 and 10 would be rather unexpected in GP sentences because of the noncanonicality of the OVS word order (see section 1.3).

Similarly to Qian et al. (2018), the relationship between response accuracy and RTs was analyzed. The motivation was to test whether the participants who answered the comprehension question incorrectly also read the sentence faster, which could be interpreted as an indication of heuristic processing. Only GP sentences were analyzed in this way, since the non-GP sentences generally yielded a low rate of incorrect answers (see section 2.2.2). Linear mixed-effects models were again used predicting residual RTs by response accuracy which was sum coded (-0.5 for correct response and +0.5 for incorrect response). Item and participant were used as random effects. The random effects structure was again determined following instructions in Bates et al. (2015).

The response accuracy was analyzed using logit-mixed models (see Jaeger, 2008). Sentence type and question type were used as fixed effects, including the interaction term between them. For these purposes, sentence type was coded using treatment contrasts (with non-GP as a baseline condition) and question type was sum coded (-0.5 for Question 6b and +0.5 for Question 6a). Participant and item were used as random effects. The random effects structure (the inclusion of random slopes) was determined following instructions in Bates et al. (2015). The beta estimates, standard errors, z-values, and p-values are reported.

2.2 Results

2.2.1 Reaction times

Raw mean RTs for each region are presented in Table 2. Figure 1 shows the transformed RTs for each word in the GP and non-GP sentences.

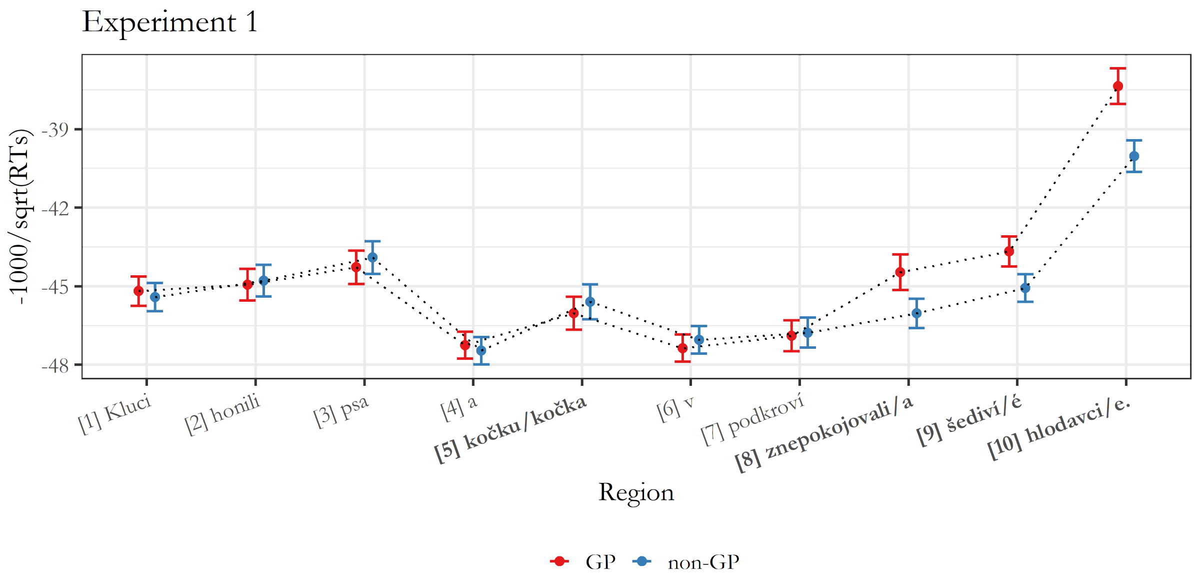
Table 2

Regions in the two conditions (GP and non-GP), raw reaction times together with the corresponding 95% confidence intervals. In the GP condition, region 5 (kočku) is the ambiguous NP, region 8 (znepokojovali) is the disambiguating verb and regions 9 and 10 represent the subject NP of the second clause.

	GP region	**GP mean RTs**	**non-GP region**	**non-GP mean RTs**
1	Kluci (Boy-NOM.M.PL)	571.45 [550.03, 592.86]	Kluci (Boy-NOM.M.PL)	555.16 [537.43, 572.89]
2	honili (chase-3PL.M.PST)	593.41 [568.98, 617.85]	honili (chase-3PL.M.PST)	591.12 [570.11, 612.12]
3	psa (dog-ACC.M.SG)	623.48 [598.26, 648.71]	psa (dog-ACC.M.SG)	635.65 [609.03, 662.28]
4	a (and)	511.67 [490.71, 532.62]	a (and)	498.87 [483.00, 514.74]
5	**kočku** (cat-ACC.F.SG)	588.48 [556.71, 620.24]	**kočka** (cat-NOM.F.SG)	623.02 [590.17, 655.88]
6	v (in)	501.99 [485.63, 518.36]	v (in)	514.75 [497.37, 532.14]
7	podkroví (attic-LOC.N.SG)	541.34 [518.52, 564.16]	podkroví (attic-LOC.N.SG)	529.96 [510.73, 549.19]
8	**znepokojovali** (worry-3PL.M.PST)	667.38 [631.50, 703.25]	**znepokojovala** (worry-3SG.F.PST)	548.71 [528.93, 568.49]
9	**šediví** (grey-NOM.M.PL)	630.40 [602.02, 658.78]	**šedivé** (grey-ACC.M.PL)	555.19 [539.26, 571.12]
10	**hlodavci.** (rodents-NOM.M.PL)	980.02 [934.03, 1026.01]	**hlodavce.** (rodents-ACC.M.PL)	790.17 [753.98, 826.37]

Figure 1

Mean transformed RTs for each word in garden-path (GP) and non-garden-path (non-GP) sentences together with their 95% confidence intervals. In the GP condition, region 5 (kočku) represents the ambiguous NP, region 8 (znepokojovali) is the disambiguating verb and regions 9 and 10 represent the subject NP of the second clause.



The linear mixed-effects models[1] yielded strong effects of sentence type for region 8 ($\beta = 1.55$, SE = 0.454, t = 3.413, $p < 0.01$), region 9 ($\beta = 1.503$, SE = 0.303, t = 4.963, $p < 0.001$) and region 10 ($\beta = 2.51$, SE = 0.46, t = 5.463, $p < 0.001$). In other words, the mean RTs for GP sentences were significantly slower than the grand mean.

2.2.2 Response accuracy

There was relatively high between-participant variability in the response accuracy. The mean accuracy was 81.18% (the median was 87.5%), with the lowest score being 45.83% and the highest score being 100%. The between-item variability was between 63.22% to 93.1% (mean 81.18% and median 82.76%).

Table 3

Number of correct and incorrect answers to the two types of comprehension questions following GP and non-GP sentences.

6a: Did the boys chase the cat?	**GP sentence**	**non-GP sentence**
correct	342	481
incorrect	180	41
% incorrect	34.48%	7.85%
6b: Did the rodents worry the cat?	**GP sentence**	**non-GP sentence**
correct	407	465
incorrect	115	57
% incorrect	22.03%	10.92%

The descriptive statistics for incorrect responses by sentence type are presented in Table 3. The logit mixed-effects model[2] revealed a significant main effect for sentence type ($\beta = 1.503$, SE = 0.173, z =

[1] All three models reported included Sentence type as random slopes for both random effects (Participants and Items).

[2] The model included Sentence type as a random slope for Participants and no random slope for Items.

8.665, $p < 0.001$), but only a marginal effect for question type (β = -0.436, SE = 0.223, t = -1.955, p = 0.051). This indicates that GP sentences had a higher proportion of incorrect responses compared to non-GP questions. However, there was no strong evidence for a difference between the two comprehension questions. Crucially, there was a significant interaction between sentence type and question type (β = 1.227, SE = 0.273, z = 4.494, $p < 0.001$). To understand this interaction, follow up analyses predicted the likelihood of an incorrect response by a simple effect of question type, with separate models being run for the two different sentence types.[3] The model for GP sentences yielded a significant effect of question type: question 6a was answered less correctly than question 6b following GP sentences (β = 0.1, SE = 0.245, z = 4.074, $p < 0.001$). However, no effect of question type was found for response accuracy on non-GP sentences.

2.2.3 Relation between response accuracy and RTs

There were small effects of response accuracy on RTs for GP sentences on regions 8[4] (β = 1.059, SE = 0.536, t = 1.976, $p < 0.05$) and 10[5] (β = 1.838, SE = 0.76, t = 2.42, $p < 0.05$). A post-hoc analysis revealed significant effects for regions 4[6] (β = -1.337, SE =0.437, t = -3.062, $p < 0.01$) and region 5[7] (β = -1.023, SE = 0.495, t = -2.068, $p < 0.05$). This shows that participants who answered incorrectly had a slight tendency to react faster on region 4 (the between-clause conjunction) and 5 (the ambiguous noun) but conversely to react slower on the disambiguating verb (region 8) and on the final region of the sentence (region 10).

2.2.4 Post-hoc analysis: plausibility of the second clause

The observed differences in response accuracy could be due to the different meanings of the second clauses in the GP and non-GP conditions. More precisely, the difference lies in the subject and object switch between the conditions – in the GP condition (5a), the meaning of the second clause is ‘grey rodents in the attic worried a cat’, whereas in the non-GP condition (5b), it is the other way round, i.e. ‘a cat in the attic worried grey rodents’. This difference could possibly influence the results because the two meanings might differ in their plausibility. Therefore, a post-hoc investigation into sentence plausibility was run to test whether the second clause differs in plausibility between the two sentence conditions.

As materials, the second clauses of the experimental sentences from Experiment 1 were used. The clauses from GP conditions were switched to the SVO word-order to eliminate any possible effect of word-order on ratings. This provided twenty-four items with two conditions. For example, (GP) *Šediví hlodavci v podkroví znepokojovali kočku* [grey rodents in the attic worried a cat] and (non-GP) *Kočka v podkroví znepokojovala šedivé hlodavce* [a cat in the attic worried grey rodents]. Every participant read only one condition of each item. Eighteen filler sentences, which were either ungrammatical or nonsensical were also included. The task was to read the sentences and evaluate their plausibility on a five-point scale ranging from 1 (completely implausible) to 5 (completely plausible). This post-hoc test was run on-line using IbexFarm 0.3.9 (http://spellout.net/ibexfarm/). The participants were recruited over our institute’s Facebook page, which provided a sample of 74 native Czech speakers (45 female, 29 male; mean age 35.4 years). The mean ratings were 4.19 for GP clauses and 4.27 for the non-GP clauses. In the linear mixed-effects model with sentence type as a fixed effect (sum coded as -

[3] Both models included Question type as a random slope for Items and no random slope for Participants.

[4] Resulting linear mixed-effects model did not include any random slopes due to singularity issues.

[5] Resulting linear mixed-effects model included response accuracy as a random slope for both Items and Participants.

[6] Resulting linear mixed-effects model did not include any random slopes due to singularity issues.

[7] Resulting linear mixed-effects model included response accuracy as a random slope for Participants and no random slope for Items.

0.5 for non-GP and +0.5 for GP) and the item and participant as random effects, these ratings were not significantly different (β = -0.08, SE = 0.091, t = -0.874, p = 0.392). We may conclude that the plausibility of the second clauses in both conditions tested in Experiment 1 was high and there was no evidence that the plausibility of the two conditions differed. The observed results for response accuracy can thus be plausibly related to the sentence ambiguity.

2.3 Discussion

Experiment 1 showed several important findings. In the reaction time analysis, there was a clear garden-path effect: the RTs on regions 8, 9, and 10 were slower in GP sentences than in non-GP sentences. The slow-down at region 8 may be interpreted as a sign of an – at least attempted – reanalysis which should take place after encountering the verb of the second clause (i.e. *znepokojovali*, 'to worry' in the example sentence). The slower RTs on following regions may be interpreted as spillover effects, i.e. continual slow-down caused by the need to reanalyze the sentence structure. Another interpretation might be that regions 9 and 10 cause more processing difficulties in cases when the parser fails to identify and repair the garden-path while processing the verb. It might be that there is an expectation of an object following the verb of the second clause due to a canonicity of the SVO word order. However, these two possibilities (that it is a spillover effect of the disambiguating verb and that it is an additional surprisal effect) cannot be distinguished based on the experimental design used.

Additionally, the analysis highlighted specific effects of response accuracy on RTs during reading: Regions 8 and 10 showed a slow-down in RTs for incorrect responses on comprehension questions that were presented after reading the GP sentences. A post-hoc analysis showed the opposite effect for RTs on regions 4 and 5, which were faster for participants who answered the comprehension questions incorrectly. It should be noted however, that these effects were rather weak, and we discuss them together with the findings of the two other experiments in the General discussion.

In the analysis of response accuracies for the comprehension questions, several effects were found. Importantly, the general pattern of the results was very similar to the findings of previous studies within the Good-Enough Approach (e.g. Christianson et al., 2001). The comprehension questions targeting the initial garden-path analysis (6a) yielded clearly more incorrect answers after the GP sentences than after the non-GP sentences. However, there was a relatively high rate of incorrect answers after question targeting the second clause (6b), with this question being answered incorrectly after GP sentence, in comparison to after non-GP sentences. In other words, GP sentences yielded generally more incorrect answers than non-GP sentences.

This is a potentially important finding in relation to the resulting representation of GP sentences. If the representation was faithful to the input, there would be no reason for a lower response accuracy for question 6b after a GP sentence. It may be that GP sentences generate more incorrect answers than control sentences do, independently of question type. This would be in accordance with the view that the resulting representation of the sentence may be disrupted due to a cognitive overload, which arises while conducting an uneasy reanalysis of a GP structure. Experiment 2 aims to test this idea by employing more types of comprehension questions using otherwise identical stimuli.

3 Experiment 2

The aim of Experiment 2 was to broaden the findings of Experiment 1, focusing on the resulting interpretation of the sentences through comprehension questions. If the GP sentences present a processing task that is demanding and which often leads to a processing failure (i.e. to a disrupted and confused final representation), it could be expected that the participants would respond incorrectly even to different comprehension questions than the two used in Experiment 1. Therefore, two more comprehension questions were added to the experimental design:

(6c) Honili kluci psa?
'Did the boys chase the dog?'

(6d) Znepokojovali hlodavci psa?
'Did the rodents worry the dog?'

Question 6c targets the correct interpretation of the first clause, which should be straightforward since the correct analysis should be done prior to noticing any problems with the syntactic analysis. For both the GP and non-GP sentences, the correct answer to question 6c is "yes". Question 6d targets an analysis which should never occur during the reading of the sentence (it is not syntactically licensed at any point during processing). For both the GP and non-GP sentences, the correct answer to question 6c is "no". The analysis of the response accuracy for these questions may give us more insights into the underlying reasons behind why there is an inability to process the GP sentences.

The experiment was preregistered on the Open Science Framework: https://osf.io/t3ecm and the data are available there too: https://osf.io/bjas8/

3.1 Method

3.1.1 Participants

Seventy-six undergraduate students from Charles University (61female, 15 male; mean age 21.9 years) participated in Experiment 2. All participants were native speakers of Czech and participated for course credit and none of them previously participated in Experiment 1.

3.1.2 Materials

The same 24 experimental items as in Experiment 1 were used. As in Experiment 1, each participant received only one condition of each item based on the Latin-square design and thus received six examples of each condition. 142 filler sentences were used, whose syntactic structure was different from the experimental sentences and which were also followed by a yes-no comprehension question. Seventy-two of these served as experimental items in two other experiments. The questions were counterbalanced, so that there was an even proportion of negative and positive correct answers in the whole experiment.

3.1.3 Procedure

Experiment 2 used the same procedure as Experiment 1. The experiment took about 25–30 minutes.

3.1.4 Data analysis

As in Experiment 1, the response accuracy on filler items was checked first. The average response accuracy was 96.1% (median 96.61%) and no participant had response accuracy for filler items that was less than 70%. No participant was therefore excluded based on the response accuracy criterion.

The reaction times were trimmed very conservatively; only those data points that were clearly discontinuous (less than 130 ms and more than 10 seconds) were excluded. This accounted for 0.1% of the data. Since the data were not normally distributed, the Box-Cox test was used to establish the ideal data transformation method. This test yielded a score of $\lambda = -0.717$, which means that the ideal transformation would be inversely transformed squared RTs (1/sqrt(RTs)). Again, the inversely transformed squared RTs were multiplied by -1000 so that the coefficients had the same sign, and to avoid very small values or overly restricted ranges for the dependent variable values.

The RT analyses were done using linear mixed-effects models the same way as in Experiment 1: (i) a linear mixed-effects model with sentence type as a fixed effect and participant and item as random effects was run on the whole data and (ii) a linear mixed-effects model with response accuracy as a fixed effect and participant and item as random effects were computed for the GP sentences. In both cases, the random-effects structure (the inclusion of random slopes) was again determined following Bates et al. (2015). The beta estimates, standard errors, t-values, and p-values are reported (only for the statistically significant results). The regions for the sentence types are presented in Table 4. Again, the target regions for the analysis were regions 8 (the disambiguating region), 9 (possible spillover region), and 10 (possible spillover region).

Response accuracy was again analyzed using logit-mixed models (see Jaeger, 2008). This time, the interaction between sentence type and question type are not reported because the model with the specified interaction term failed to converge. Thus, the model included sentence type and question type as fixed effects and participant and items as random effects. The random effects structure (the inclusion of random slopes) was determined following Bates et al. (2015). The beta estimates, standard errors, z-values, and p-values are reported. In the model, sentence type was coded using treatment contrast (with non-GP as the baseline condition). Question type was coded using repeated contrasts (Schad, Vasishth, Hohenstein, & Kliegl, 2020), shown in Table 4. The reason to use repeated contrast coding was to allow for comparisons between Question 6a to Question 6b, Question 6b to Question 6d and Question 6d to Question 6c.

Table 4: Repeated contrast coding matrix for question type used in the response accuracy analysis.

	6a-6b	6b-6d	6d-6c
Question 6a	-0.75	-0.5	-0.25
Question 6b	0.25	-0.5	-0.25
Question 6c	0.25	0.5	0.75
Question 6d	0.25	0.5	-0.25

Because it was not possible to analyze the interaction between sentence type and question type (models containing this interaction did not converge), separate analyses for the effect of sentence type for each question were conducted.

3.2 Results

3.2.1 Reaction times

Raw mean RTs for each region are presented in Table 5. Figure 2 shows the transformed RTs for each word in the GP and non-GP sentences.

Table 5

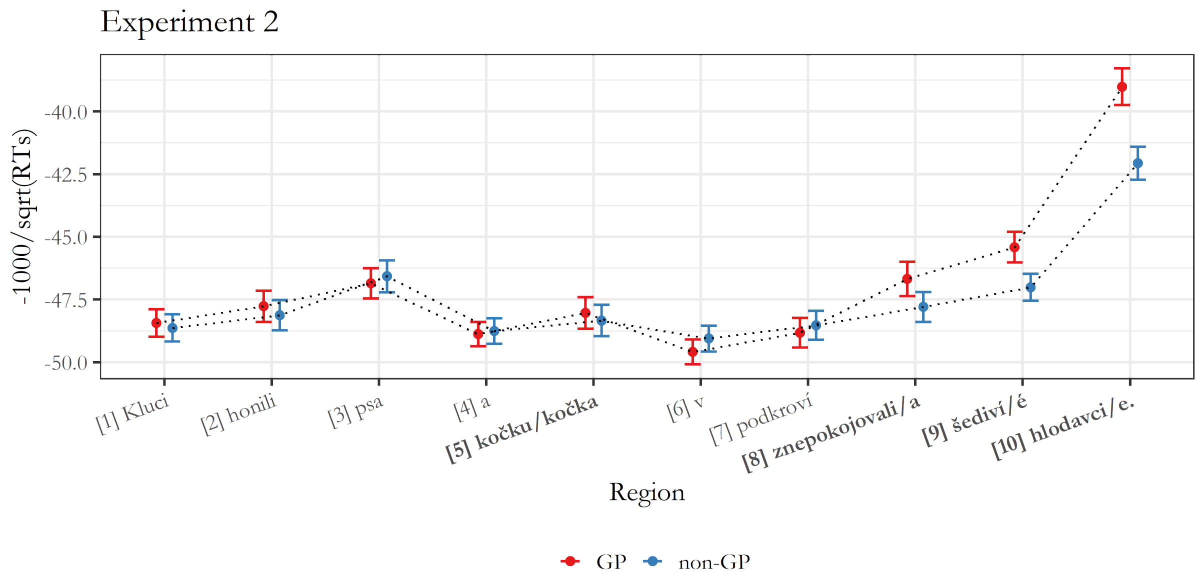
Regions in the two conditions (GP and non-GP), raw reaction times, together with the corresponding 95% confidence intervals. In the GP condition, region 5 (kočku) represents the ambiguous NP, region 8 (znepokojovali) is the disambiguating verb and regions 9 and 10 represent the subject NP of the second clause.

	GP region	**GP mean RTs**	**non-GP region**	**non-GP mean RTs**
1	Kluci (Boy-NOM.M.PL)	478.16 [461.88, 494.43]	Kluci (Boy-NOM.M.PL)	478.12 [458.38, 497.86]
2	honili (chase-3PL.M.PST)	515.77 [493.88, 537.67]	honili (chase-3PL.M.PST)	500.98 [480.34, 521.63]
3	psa (dog-ACC.M.SG)	527.57 [507.47, 547.67]	psa (dog-ACC.M.SG)	554.28 [527.18, 581.39]
4	a (and)	459.61 [439.84, 479.38]	a (and)	466.34 [451, 481.68]
5	**kočku** (cat-ACC.F.SG)	508.66 [487.89, 529.43]	**kočka** (cat-NOM.F.SG)	504.86 [481.8, 527.92]
6	v (in)	442.1 [430.12, 454.08]	v (in)	463.78 [447.48, 480.09]

7	podkroví (attic-LOC.N.SG)	481.72 [462.75, 500.69]	podkroví (attic-LOC.N.SG)	479.98 [463.88, 496.09]
8	**znepokojovali** (worry-3PL.M.PST)	571.72 [543.02, 600.42]	**znepokojovala** (worry-3SG.F.PST)	506.03 [486.46, 525.6]
9	**šediví** (grey-NOM.M.PL)	569.22 [545.66, 592.78]	**šedivé** (grey-ACC.M.PL)	503.36 [488.97, 517.75]
10	**hlodavci.** (rodents-NOM.M.PL)	910.57 [859.21, 961.94]	**hlodavce.** (rodents-ACC.M.PL)	706.39 [673.45, 739.33]

Figure 2

Mean transformed RTs for each word in garden-path (GP) and non-garden-path (non-GP) sentences together with their 95% confidence intervals in Experiment 2. In the GP condition, region 5 (kočku) represents the ambiguous NP, region 8 (znepokojovali) is the disambiguating verb and regions 9 and 10 represent the subject NP of the second clause.



The results were very similar to those of Experiment 1. Significant effects of sentence type were found for region 8 (β = 1.188, SE = 0.38, t = 3.128, p < 0.01), region 9[8] (β = 1.81, SE = 0.35, t = 5.17, p < 0.001) and region 10[9] (β = 3.057, SE = 0. 584, t = 5.239, p < 0.001).

3.2.2 Response accuracy

Participants' response accuracies for the experimental sentences were 54.17%–100% with a mean accuracy of 79.11% (median 79.17%). The accuracies for items ranged from 69.74%–90.79% with a mean accuracy of 79.11% (median 79.61%).

Table 6

Number of correct and incorrect answers on four types of comprehension questions following GP and non-GP sentences in Experiment 2.

6a: Did the boys chase the cat?	**GP sentence**	**non-GP sentence**
correct	107	198

[8] Linear mixed-effects models for regions 8 and 9 included Sentence type as a random slope for Participants and no random slope for Items.

[9] The model included Sentence type as a random slope for both Participants and Items.

incorrect	121	30
% incorrect	53.07%	13.16%
6b: Did the rodents worry the cat?	**GP sentence**	**non-GP sentence**
correct	132	210
incorrect	96	18
% incorrect	42.11%	7.89%
6c: Did the boys chase the dog?	**GP sentence**	**non-GP sentence**
correct	202	217
incorrect	26	11
% incorrect	11.4%	4.82%
6d: Did the rodents worry the dog?	**GP sentence**	**non-GP sentence**
correct	167	210
incorrect	61	18
% incorrect	26.75%	7.89%

The descriptive statistics for response accuracy and sentence type are presented in Table 6. The logit-mixed model[10] yielded significant effects for sentence type ($\beta = 2.162$, SE = 0.206, $z = 10.479$, $p < 0.001$) and for all three question type factors: 6a-6b ($\beta = -0.519$, SE = 0.17, $z = -3.052$, $p < 0.01$), 6b-6d ($\beta = -0.611$, SE = 0.187, $z = -3.277$, $p < 0.01$), and 6d-6c ($\beta = -0.99$, SE = 0.229, $z = -4.329$, $p < 0.001$). Thus, the results indicate that more errors were produced when answering comprehension questions after GP sentences compared with after non-GP sentences. It also shows evidence that response accuracy differed between question types (Question 6a yielded significantly more errors than Question 6b; Question 6b yielded more errors than Question 6d, and Question 6d yielded more errors than Question 6c).

The separate analyses showed that the rate of incorrect answers was higher after reading GP sentences for four question types:[11] Question 6a ($\beta = 2.83$, SE = 0.349, $z = 8.117$, $p < 0.001$), Question 6b ($\beta = 3.811$, SE = 0.868, $z = 4.389$, $p < 0.001$), Question 6c ($\beta = 1.09$, SE = 0.5, $z = 2.181$, $p < 0.05$), and Question 6d ($\beta = 1.781$, SE = 0.445, $z = 4.006$, $p < 0.001$).

3.2.3 Relation between response accuracy and RTs

There were no effects of response accuracy on RTs for GP sentences. Neither did a post-hoc analysis of the rest of the regions show any effects of sentence type.

3.3 Discussion

As in Experiment 1, GP sentences in Experiment 2 showed clear garden-path effects, which is highlighted through the higher RTs on regions 8, 9, and 10. In contrast with Experiment 1, no effect was found on region 5 and there was no effect of response accuracy to the comprehension questions. Therefore, it seems that those who answered incorrectly, did not react faster on any region during the reading of the sentences.

The analysis of response accuracy yielded noteworthy findings. As in Experiment 1, comprehension questions presented after GP sentences yielded more incorrect answers than after control sentences. Importantly, the rate of incorrect answers after GP sentences was significantly higher for every

[10] Model included Sentence type as a random slope for Participants and no random slope for Items.

[11] The model for Question 6a did not contain any random slopes due to singularity issues. Models for Question 6b, Question 6c, and Question 6d included Sentence type as a random slope for Items and no random slope for Participants.

comprehension question, including Question 6c targeting the first clause (‘Boys chased a dog’) and Question 6d targeting an analysis that could not emerge during reading since it is not syntactically licensed at any point of processing. The incorrect answers to Questions 6c and 6d cannot be explained by a lingering representation, since Question 6c targets the correct analysis of the first clause which is virtually unambiguous, and Question 6d targets an analysis which should not emerge at all during reading. These findings thus suggest that readers sometimes simply failed in processing the GP sentences and that the resulting representation was disrupted and confused.

4 Experiment 3

Both Experiments 1 and 2 showed a significant tendency to respond incorrectly on comprehension questions following GP sentences, compared to non-GP sentences. Nevertheless, the tendency was even higher for questions which target the initial misanalysis. Experiment 3 tested whether this tendency is affected by the plausibility of the initial misanalysis. The plausibility of the initial misanalysis was manipulated by using inanimate nouns which were highly incompatible with the verb of the first clause.

Based on prior studies on the effects of plausibility in garden-path processing (e.g. Pickering & Traxler, 1998; Clifton, Traxler, Mohamed, Williams, Morris, & Rayner, 2003; Slattery et al., 2013; den Ouden, Dickey, Anderson, & Christianson, 2016), we may expect that even if the initial misanalysis was implausible, the readers would still follow the garden-path. The crucial difference between the plausible and implausible sentences should lie in the easier reanalysis for the implausible sentences (see Clifton et al., 2003, pp. 328–329) and also in the possibility that the readers will start to reanalyze prior to encountering the disambiguating word (i.e. the verb of the second clause; see Pickering & Traxler, 1998, p. 956). In other words, implausible GP sentences should be somehow harder to process than control sentences, but the readers should arrive at a correct analysis much more easily than for plausible GP sentences.

Based on the assumption that it should be easier to conduct the reanalysis for the implausible GP sentences, and thus arrive at a correct sentence analysis, we may predict that: (a) the rate of incorrect answers for questions targeting the initial misanalysis should be lower for implausible than for plausible GP sentences (it should also be similar to the rate of incorrect answers for non-GP sentences); (b) questions targeting the correct analysis of the first clause (like 6c) and questions targeting the correct analysis of the second clause (like 6b) should both yield less incorrect answers after implausible than after plausible GP sentences (and probably a very similar rate of incorrect answers to control sentences). But since the implausible GP sentence is still harder to process than the control sentence (because of the local ambiguity), we may tentatively predict that (c) a question targeting the analysis which is not syntactically licensed at any point in time (like 6d) may yield a higher rate of incorrect answers than the similar question for the control sentences.

The experiment was preregistered on the Open Science Framework: https://osf.io/q2bd5 and the data are freely available here: https://osf.io/bjas8/

4.1 Method

4.1.1 Participants

Ninety-four students from Charles University (81 female, 13 male; mean age = 21.19 years) participated in this experiment (part of them for course credit, part for a fee of 200 CZK). None of the participants participated in previous experiments and all participants were native speakers of Czech.

4.1.2 Materials

Twenty-four experimental items were used, and each item consisted of twelve conditions (3x4 factorial design) which were based on the two independent variables: sentence type (plausible GP, non-GP, and implausible GP, see 7a–7c) and comprehension question type (the same four question types as in Experiment 2, see 8a–8d). The same sentences as in Experiments 1 and 2 were used, but they were slightly modified because of the alignment with the implausible conditions (see Supplementary Materials for the whole list with English translation).

(7a) plausible GP condition

Kluci	honili	psa	a	kočk-u	v	podkroví
Boy-NOM.M.PL	chase-3PL.M.PST	dog-ACC.M.SG	and	cat-ACC.F.SG	in	attic

pozorovali	šediví	hlodavci.
observe-3PL.M.PST	grey-NOM.M.PL	rodents-NOM.M.PL

‘Boys chased a dog and grey rodents in the attic observed a cat.’

(7b) non-GP condition

Kluci	honili	psa	a	kočk-a	v	podkroví
Boy-NOM.M.PL	chase-3PL.M.PST	dog-ACC.M.SG	and	cat- NOM.F.SG	in	attic

pozorovala	šedivé	hlodavce.
observe-3SG.F.PST	grey-ACC.M.PL	rodents-ACC.M.PL

‘Boys chased a dog and a cat in the attic observed grey rodents.’

(7c) implausible GP condition

Kluci	honili	psa	a	bedn-u	v	podkroví
Boy-NOM.M.PL	chase-3PL.M.PST	dog-ACC.M.SG	and	box- ACC.F.SG	in	attic

pozorovali	šediví	hlodavci.
observe-3SG.F.PST	grey-ACC.M.PL	rodents-ACC.M.PL

‘Boys chased a dog and grey rodents in the attic observed a box.’

(8a) Honili kluci kočku/bednu?
‘Did the boys chase the cat/box?’

(8b) Pozorovali hlodavci kočku/bednu?
‘Did the rodents observe the cat/box?’

(8c) Honili kluci psa?
‘Did the boys chase the dog?’

(8d) Pozorovali hlodavci psa?
‘Did the rodents observe the dog?’

Similarly to Experiment 2, the correct answer to questions 8a and 8d is “no”, and the correct answer to question 8c is “yes”. For question 8b, the correct answer differs depending on the previous sentence. For plausible and implausible GP sentences, it is “yes”, for non-GP sentences, it is “no”.

It should be noted that the implausibility of the 7c condition stems primarily from the implausibility of the NP *bednu* (‘a box’, accusative) to be an object of the verb *honili* (‘(they) chased’). In other words, the implausible GP condition used in this experiment is what den Ouden et al. (2016) call “GP with early closure”. However, another possibility is that the initial NP–NP coordination between the first clause object and the ambiguous NP, such as *psa a bednu* (‘a dog and a box’), may be implausible by itself. This is possible, but it seems more likely that the plausibility of this NP–NP coordination is governed by the verb (it seems perfectly plausible to “spot a dog and a box” or to “draw a dog and a box”, but it does not seem plausible to “chase a dog and a box”, probably because it would be strange to chase a box, based on real-world knowledge). Nevertheless, even if this coordination was implausible per se, it should only deepen the implausibility of the GP structure because the implausibility of linking the NP *bednu* to the verb *honili* would go hand in hand with the implausibility of having coordinated objects *psa* and *bednu*.

Each participant received only one condition of each item based on the Latin-square design and thus received two examples of each condition. There were also 142 filler sentences, whose syntactic structures were different from the experimental sentences and which were also followed by a yes-no comprehension question. Seventy-two of these served as experimental items in two other experiments. The questions were counterbalanced so that there was an even proportion of the negative and positive correct answers in the whole experiment.

4.1.3 Procedure

The experiment took about 30 minutes and employed the same general procedure as Experiments 1 and 2.

4.1.4 Data analysis

The mean response accuracy on filler items was 96.27% (median 96.61%) and no participant had an accuracy lower than 70%. Therefore, no participants were excluded from the data analysis.

The RTs were trimmed very conservatively; only those data points that were clearly discontinuous (less than 130 ms or more than 10 seconds) were excluded. This represented 0.15% of the data. Based on the Box-Cox test (λ = -0.696), the RTs were inversely transformed squared RTs (1/sqrt(RTs)) and then multiplied by -1000 as in Experiments 1 and 2.

The RT analyses used linear mixed-effects models in the same way as Experiment 2. The difference was that sentence type is a three-level factor in Experiment 3. It was sum coded following Schad et al. (2020), which allowed for (i) comparison of the RTs of plausible GP sentences to the grand mean (Comparison A) and (ii) comparison of the RTs of implausible GP sentences to the grand mean (Comparison B). The coding matrix is shown in Table 7.

Table 7: Sum contrast coding matrix for sentence type in Experiment 3.

	Comparison A	Comparison B
Plausible GP	0.5	0
non-GP	-0.5	-0.5
Implausible GP	0	0.5

The regions for the sentence types are presented in Table 8. The target regions for the analysis of plausible GP and non-GP sentences were regions 8 (the disambiguating region), 9 (a possible spillover region), and 10 (a possible spillover region). For the analysis of implausible GP sentences, the target regions were also region 5 (the implausible NP), region 6 (possible spillover region for implausibility effects) and region 7 (possible spillover region for implausibility effects) because we may expect an early garden-path repair due to implausibility (cf. den Ouden et al., 2016).

Response accuracy was again analyzed using logit-mixed models (see Jaeger, 2008). The logit-mixed model containing sentence type and three question type factors (resulting from repeated contrast coding as a fixed effect) failed to converge. Therefore, two separate models were run: one with sentence type as a fixed effect and another with the three question type fixed effects. Both of these models contained item and participant as random effects. The random-effects structure (the inclusion of random slopes) was again determined following Bates et al. (2015). As in the Experiment 2, separate analyses for the effect of sentence type for each question were also conducted.

4.2 Results

4.2.1 Reaction times

Raw mean RTs for each region are presented in Table 8. Figure 3 shows the transformed RTs for each word in the GP and non-GP sentences.

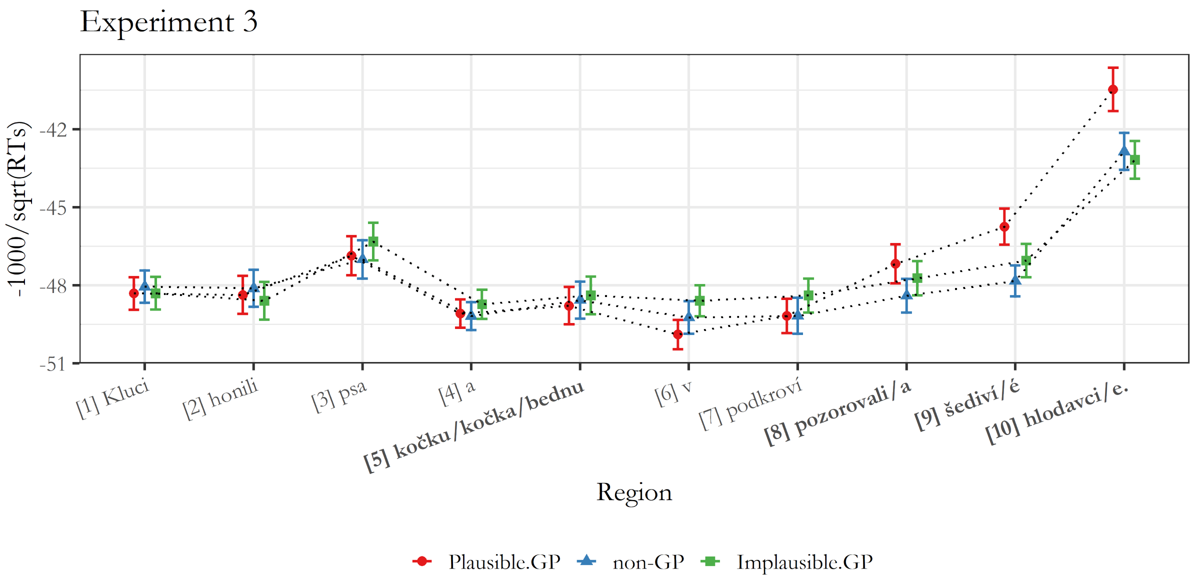
Table 8

Regions in the three sentence conditions of Experiment 3 (plausible GP, non-GP, and implausible GP), mean raw RTs together with the corresponding 95% confidence intervals. In the GP conditions, region 5 (kočku/bednu) represents the ambiguous NP, region 8 (pozorovali) is the disambiguating verb and regions 9 and 10 represent the subject NP of the second clause.

	plausible GP region	**plausible GP mean RTs**	**non-GP region**	**non-GP mean RTs**	**implausible GP region**	**implausible GP mean RTs**
1	Kluci (Boy-NOM.M.PL)	488.04 [468.19, 507.89]	Kluci (Boy-NOM.M.PL)	490.58 [470.92, 510.25]	Kluci (Boy-NOM.M.PL)	486.54 [467.34, 505.75]
2	honili (chase-3PL.M.PST)	513.31 [484.93, 541.69]	honili (chase-3PL.M.PST)	513.58 [487.42, 539.74]	honili (chase-3PL.M.PST)	505.24 [480.11, 530.36]
3	psa (dog-ACC.M.SG)	564.47 [532.89, 596.05]	psa (dog-ACC.M.SG)	564.86 [526.87, 602.84]	psa (dog-ACC.M.SG)	567.28 [537.53, 597.03]
4	a (and)	457.99 [439.35, 476.63]	a (and)	447.9 [435.48, 460.32]	a (and)	469.29 [450.06, 488.51]
5	**kočku** (cat-ACC.F.SG)	505.98 [479.19, 532.78]	**kočka** (cat-NOM.F.SG)	508.14 [480.01, 536.28]	**bednu** (box-ACC.F.SG)	517.63 [489.37, 545.88]
6	v (in)	448.07 [426.25, 469.88]	v (in)	465.01 [447.01, 483.01]	v (in)	473.91 [456.43, 491.38]
7	podkroví (attic-LOC.N.SG)	478.03 [457.13, 498.93]	podkroví (attic-LOC.N.SG)	479.41 [458.12, 500.71]	podkroví (attic-LOC.N.SG)	488.76 [469.8, 507.71]
8	**pozorovali** (watch-3PL.M.PST)	565.14 [530.19, 600.09]	**pozorovala** (watch-3SG.F.PST)	484.06 [465.93, 502.2]	**pozorovali** (watch-3PL.M.PST)	507.9 [486.43, 529.37]
9	**šediví** (grey-NOM.M.PL)	569.24 [543.46, 595.01]	**šedivé** (grey-ACC.M.PL)	484.76 [469.06, 500.46]	**šediví** (grey-NOM.M.PL)	516.27 [496.55, 536]
10	**hlodavci.** (rodents-NOM.M.PL)	884.45 [818.56, 950.34]	**hlodavce.** (rodents-ACC.M.PL)	666.83 [636.14, 697.52]	**hlodavci.** (rodents-NOM.M.PL)	670.44 [634.6, 706.27]

Figure 3

Mean transformed RTs for each word in plausible garden-path (plausible GP), non-garden-path (non-GP) and implausible garden-path (implausible GP) sentences together with their 95% confidence intervals in Experiment 3.



Based on the linear mixed-effects models,[12] a significant effect for Comparison A was found for region 8 (β = 1.279, SE = 0.492, t = 2.598, p < 0.05), region 9 (β = 2.282, SE = 0.482, t = 4.731, p < 0.001), and region 10 (β = 3.309, SE = 0.646, t = 5.121, p < 0.001). This indicates that the RTs for plausible GP sentences on these regions were slower than the grand mean. For the Comparison B, there were significant effects for region 6[13] (β = 1.323, SE = 0.559, t = 2.366, p < 0.05), region 7 (β = 1.091, SE = 0.477, t = 2.288, p < 0.05), and region 10 (β = -1.906, SE = 0.445, t = -4.278, p < 0.001). This indicates that the RTs for implausible GP sentences were slower than the grand mean for regions 6 and 7, but faster for the sentence final region 10.

4.2.2 Response accuracy

Participants' response accuracies for the experimental sentences ranged from 54.17% to 100% and the mean score was 84.13% (the median was 87.5%). The range between items was 75.53%–92.55%, the mean score was 84.13% (the median was 84.04%).

Table 9

Number of correct and incorrect answers on four types of comprehension questions following plausible GP, non-GP, and implausible GP sentences in Experiment 3.

8a: Did the boys chase the cat/box?	**plausible GP**	non-GP	**implausible GP**
correct	80	160	164
incorrect	108	28	24
% incorrect	57.45%	14.89%	12.77%
8b: Did the rodents observe the cat/box?	**plausible GP**	non-GP	**implausible GP**
correct	126	169	176

[12] The models reported for regions 8, 9, and 10 included Comparison A as a random slope for both Items and Participants.

[13] The model for regions 6 and 7 included Comparison B as a random slope for both Items and Participants.

incorrect	62	19	12
% incorrect	32.98%	6.38%	10.11%
8c: Did the boys chase the dog?	**plausible GP**	non-GP	**implausible GP**
correct	170	181	178
incorrect	18	7	10
% incorrect	9.57%	3.72%	5.32%
8d: Did the rodents observe the dog?	**plausible GP**	non-GP	**implausible GP**
correct	148	181	165
incorrect	40	7	23
% incorrect	21.28%	3.72%	12.23%

The descriptive statistics for response accuracy and sentence type are presented in Table 9. In the model which used sentence type as the only fixed effect,[14] both plausible and implausible GP sentence types were significantly different from non-GP sentences: plausible GP (β = 1.856, SE = 0.167, z = 11.146, p < 0.001) and implausible GP (β = 0.383, SE = 0.187, z = 2.046, p < 0.05). The second model which used three question type factors resulting from repeated contrast coding[15] showed significant effects for all factors: 6a-6b (β = -0.75, SE = 0.162, z = -4.621, p < 0.01), 6b-6d (β = -0.361, SE = 0.174, z = -2.079, p < 0.05), and 6d-6c (β = -0.789, SE = 0.217, z = -3.642, p < 0.001).

As in Experiment 2, the separate analyses showed that for each question type the rate of incorrect answers was higher after plausible GP sentences than after non-GP sentences:[16] Question 8a (β = 2.522, SE = 0.319, z = 7.909, p < 0.001), Question 8b (β = 2.169, SE = 0.551, z = 3.937, p < 0.001), Question 8c (β = 1.056, SE = 0.47, z = 2.244, p < 0.05), and Question 8d (β = 2.236, SE = 0.466, z = 4.8, p < 0.001). The implausible GP sentences yielded a different pattern, whereby the only significant effect was for Question 8d: this question was answered incorrectly more often after implausible GP sentences than after non-GP sentences (β = 1.389, SE = 0.471, z = 2.95, p < 0.05). Importantly, Question 8a which targeted the initial garden-path misanalysis did not reveal a significant effect.

4.2.3 Relation between response accuracy and RTs

As in Experiment 2, the linear mixed-effects modelling did not show any effects of response accuracy on RTs. Thus, there was no significant difference in RTs between GP sentences which had incorrect responses to the comprehension questions and those which were answered correctly.

4.3 Discussion

Experiment 3 fully replicated the findings of Experiment 2 in terms of the differences in RTs and response accuracy between plausible GP and non-GP sentences. It was found that regions 8, 9, and 10 yielded longer RTs for plausible GP sentences than for non-GP sentences (and that there were no other significant differences in RTs between these sentences). Also, it was found that plausible GP sentences yielded more incorrect answers for each question type than non-GP sentences. Also, no effect of response accuracy on RTs was found (as in Experiment 2, but partially in contrast with Experiment 1).

The implausible GP sentences had a slightly different pattern for RTs. There were significant slow-down effects in regions 6 and 7. This can be interpreted as evidence that the ambiguous word (e.g. *bednu* 'a

[14] The model did not include any random slope due to singularity issues.

[15] The model included Comparison 6a-6b as a random slope for Items. No other random slopes were used due to singularity issues.

[16] The models run in these separate analyses had no random slopes (due to convergence problems) with the exception of the model for Question 8b which contained Sentence type as random slope for Items.

box', accusative) followed a garden-path, but a reanalysis was triggered sooner than for plausible GP sentences because of the implausibility of processing this word as the object of the first clause (it seems strange to chase a dog and a box at the same time). There was also a significant effect on region 10, which showed that participants' RTs were faster than the grand mean in this region while reading implausible GP sentences. This indicates that the grand mean is elevated by the RTs for plausible GP conditions (mean 884.45 ms) which are on average more than 200 ms longer than both non-GP conditions (mean 666.83 ms) and the implausible GP condition (mean 670.44 ms), see Table 7.

In sum, the garden-path effect can be observed for implausible GP sentences too, but it had a weaker impact on sentence processing because of the general implausibility of the initial analysis. This is precisely what was predicted based on the previous literature (e.g. Pickering & Traxler, 1998; Clifton et al., 2003; den Ouden et al., 2016).

The analysis of response accuracy for implausible GP sentences also showed significant differences based on the question types. First, and most importantly, the response accuracy for Question 8a (targeting the initial misanalysis) was particularly high and the rate of incorrect answers for this question after implausible GP sentences was not significantly different from non-GP sentences. This could indicate that the recovery from a GP sentence (which is needed for correctly answering this question) was facilitated by the plausibility of the initial misanalysis. However, the lower rate of incorrect answers for this question might also stem from the wording of the comprehension question, because it seems implausible by itself to answer "yes" to a comprehension question *Did the boys chase a box?* We will return to this issue in the General discussion.

Interestingly, the only comprehension question which had a higher rate of incorrect answers after implausible GP sentences, compared to after non-GP sentences was Question 8d, which targeted an interpretation that could not emerge during the processing of the sentence. This effect was not as strong as for the difference between plausible GP sentences and non-GP sentences, but it was still clear. We may say that implausible GP sentences present an easier task for interpretation than plausible GP sentences, but they still yield slightly more incorrect answers than the control sentences.

5 General discussion

The aim of this paper was to address the two limitations of previous research on the Good-Enough processing of garden-path (GP) sentences, namely to examine a different, more difficult type of GP structure, and to target the resulting representation of the sentence using several comprehension question types.

A similar pattern as in all previous studies of Good-Enough processing of GP sentences arose in all three self-paced reading experiments: the questions targeting the initial misanalysis (e.g. *Did the boys chase the cat?*) yielded a significantly higher rate of incorrect answers for the GP sentences than for the non-GP controls.

However, the experiments described in this paper document a general tendency towards a lower response accuracy for the three other comprehension questions after GP sentences than after the controls, this was the case for (i) questions targeting the correct analysis of the second clause (e.g. *Did the rodents worry the cat?*), (ii) questions targeting the correct analysis of the first clause (e.g. *Did the boys chase a dog?*), and (iii) questions targeting an analysis which was not licensed by the syntax at any point of the sentence processing (e.g. *Did the rodents worry the dog?*). All three questions yielded a consistently lower response accuracy when following GP sentences than when following the non-GP controls. This highlights that the resulting representations of the GP sentences were often disrupted, possibly due to cognitive overload, arising when conducting an uneasy reanalysis of the given GP structure. If the resulting sentence representation was faithful to the input, there would be no reason for a lower response accuracy for these questions after the GP sentences.

Importantly, the question targeting the meaning of the second clause (e.g. *Did the rodents worry the cat?*), yielded a high rate of incorrect answers for GP sentences in all three experiment (22.03% in Experiment 1, 42.11% in Experiment 2, and 32.98% in Experiment 3). The correct answer to this question requires a successful reanalysis, which must follow three steps: (i) the link between NP *kočku* ('a cat', accusative) and first-clause verb *honili* ('(they) chased') must be identified, (ii) it must be detached, and (iii) the NP *kočku* must be attached to the verb *znepokojovali* ('(they) worried') as its object (cf. Van Dyke & Lewis, 2003). The higher rate of incorrect answers for this question after GP sentences can thus be interpreted as a sign that the reanalysis of the original misparsing was not successful. This contrasts with Slattery et al. (2013) who argue that readers conduct a full and complete reanalysis while reading garden-path sentences. There are multiple explanations for this discrepancy.

One possible explanation is a general crosslinguistic difference between relatively free word order languages such as Czech and rather fixed word order languages such as English considering garden-path sentence processing. Kiel Christianson (personal communication) suggests that languages with a less fixed word order may not allow the readers/hearers to easily chunk the input into substrings, using their own phrasal interpretations. Instead, the input may be chunked into certain lexical groupings that might be noncontiguous in the input stream and the relationship between these groupings may be fragile. This could be the reason why the GP sentences, in relatively free word order languages, may be susceptible to more misanalyses than in languages with a rather fixed word order. This idea is tempting and warrants testing in future research on garden-path processing.

Another possibility is that the contradiction stems from the differences between the analyzed GP sentences. A similar type of GP sentence to the one examined in the current paper is the noun phrase/sentence coordination ambiguity structure (e.g. *The publisher called up the editor and the author refused to change the book's ending* - analyzed by Christianson & Luke, 2011). Strikingly, this type of ambiguity yielded only a very low rate of incorrect answers in English. The difference here, however, is that in English the second clause with an active verb is necessarily in SVO word order and the second clause verb, therefore, almost instantly signals to the parser that the ambiguous NP (*the author*) is a subject of the second clause (and not the object of the first clause). Unlike in English, in the Czech GP sentence *Kluci honili psa a kočku v podkroví znepokojovali šediví hlodavci*, the second clause has an OVS word order and the ambiguous NP (*kočku*) is an object of the second clause. Crucially, OVS word order is possible in Czech, but it is not a canonical word order, it is marked and is typically used to focus the subject (Jasinskaja & Šimík, in press; Siewierska & Uhlířová, 1998). Thus, the OVS word order may be rather unexpected in the second clause of the sentence. Therefore, the higher rate of incorrect answers for the question targeting the second clause in the current paper may be due to a highly demanding reanalysis, resulting from the OVS word order.

Using the apt terminology of a diagnosis model by Fodor & Inoue (1994), the overt symptom (verb of the second clause) is not very informative about the nature of the error (the ambiguous NP). Since SVO is a canonical word order in Czech, the parser would probably start looking for a subject once it encounters the second clause verb (and not for an object). However, there is no NP available in the previous input, which would stand for a subject since the ambiguous NP is in accusative case. Moreover, the OVS word order is probably less preferred, than the NP coordination between the first clause object and the ambiguous NP. Therefore, the parser may, at least in some cases, not even consider linking the second clause verb with the ambiguous NP as its object while reading the second clause verb, and may proceed further with the "attach anyway" strategy (see Fodor & Inoue, 1998). The following two regions (an adjective and a noun in nominative case) represent another symptom, which could lead the parser to a correct interpretation (that the ambiguous NP is in fact an object of the second clause). However, given the previous parsing, it may be by itself unexpected and may cause even more processing disruption. The rationale is that the parser already resigned its search for the subject and expected an object to follow the verb because it just expects the canonical SVO word order. This is supported by the fact that the GP sentences yielded significant effects on the RTs in the two regions following the second clause verb (i.e. regions 9 and 10). With the present design, one cannot distinguish such an effect of an unexpected subject from a spillover effect of the disambiguating region occurring at regions 9 and 10.

In other words, the pattern of results found in the current study, may be due to the type of GP sentence used, which are quite different from the GP sentences used in previous studies. This stresses the importance of examining more types of GP structures in different languages and with varying levels of processing difficulty from the perspective of Good-Enough processing. A comparison between the present results and results of previous studies suggests a range of difficulty levels of GP structures, which is consequently related to a range of outcome representations. Some GP structures (such as *The publisher called up the editor and the author refused to change the book's ending*, as examined by Christianson & Luke, 2011) seem to be easily repaired, at least based on a very high response accuracy for comprehension questions targeting initial misanalyses, as well as no clear differences in response accuracy for GP sentences and controls. Other GP structures (such as *While Anna dressed the baby played in the crib*, examined by Christianson et al., 2001) are apparently harder and result in a lingering misanalysis. It may be that, apart from the lingering misanalysis, the resulting representation of such GP sentences is full and faithful to the input (as argued by Slattery et al., 2013). However, there may be even harder GP structures (such as the one examined in this paper), which are extremely demanding to process to the full extent. The present results suggest that readers may often end up with a disrupted and incoherent representation of these sentences (see also Fodor & Inoue, 1994, 1998).

This interpretation of the results in the present study is, similarly to previous studies, based on the assumption that the answers to comprehension questions reflect the sentence representation accurately. However, there are various factors at play, which may be influencing the response accuracy. Christianson & Luke (2011) convincingly show how specific comprehension questions can bias readers towards incorrect answers for non-GP sentences. Two factors seem to be potentially important, namely response plausibility and acquiescence.

The plausibility of the possible answers (based on participants' real-world knowledge) may cause systematic differences between the response accuracies for different questions. For example, participants may be generally inclined to answer "no" to questions such as *Did the boys chase a box?* (used in Experiment 3 for implausible GP sentences), but "yes" to questions such as *Did the boys chase a cat?* (used in all three experiments). We may assume that the role of plausibility would be rather major in cases where the readers fail to form a coherent and full representation of the sentence. Thus, plausibility may be one of the explanations for the low response accuracies on questions targeting the initial misanalysis.

Another possibly important factor may be acquiescence bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). One would expect participants to have a higher general tendency to answer "yes" than "no", especially if they failed to form a coherent representation of the sentence and would thus be unsure about the correct answer. Acquiescence would cause more incorrect answers for questions 6a (targeting the initial misanalysis, e.g. *Did the boys chase a cat?*) and 6d (targeting an analysis that is not syntactically licensed at any point of processing, e.g. *Did the rodents worry the dog?*), because the correct answer to these questions is "no". On the contrary, this bias would cause higher response accuracies for question 6c (targeting the correct interpretation of the first clause) because the correct answer to this question is "yes". Moreover, acquiescence may cause differences in response accuracy for question 6b (targeting the correct analysis of the second clause) for GP and non-GP sentences because the correct answer for GP sentences is "no", but it is "yes" for non-GP sentences. A possible indirect measure of the effect of acquiescence would be the response accuracy for filler items since they were balanced with respect to the correct answer (half of them required a "yes" answer and half a "no" answer). Based on the fact that response accuracies were very high for filler items in all experiments (over 93%), we may assume that acquiescence only had a small effect on the results. This is important for the interpretation of the results concerning question 6b, where all three experiments yielded a striking difference between GP and non-GP sentences (Exp1: 22.03% vs. 10.92%, Exp2: 42.11% vs. 7.89%, Exp3: 32.98% vs. 6.38%). This difference is hard to account for solely by acquiescence, considering the response accuracy for filler items.

The results indicate that the questions differed in their accuracy rate. Question 6a (targeting the initial misanalysis) yielded the lowest response accuracy overall, and question 6b (targeting the correct analysis of the second clause) yielded significantly more incorrect answers than question 6d (targeting the syntactically unlicensed analysis), which in turn had more incorrect answers than question 6c (targeting the correct analysis of the first clause). The presence of factors, such as response plausibility, which may have influenced participants' responses makes it difficult to interpret these differences clearly. Still, the fact that all four questions yielded significantly more incorrect answers after GP sentences than after the control non-GP sentences suggests that there is still a reliable effect of sentence type (especially if we do not expect the acquiescence to play a major role in answering question 6b).

The main finding of this paper, namely that readers often fail in processing certain GP structures, can be explained by cognitive overload, demonstrated by the processing difficulties that occur during the repair process. This is in accordance with other studies suggesting that the ease of processing syntactic ambiguities is related to working memory capacity (Christianson et al., 2006; Engelhardt, Nigg, Carr, & Ferreira, 2008; Stella & Engelhardt, 2019) or general intelligence (Engelhardt et al., 2017). Importantly, there is evidence from such studies (see Engelhardt et al., 2017; Stella & Engelhardt, 2019) that processing speed is correlated with ambiguity resolution – faster reading of the sentence was related to higher response accuracy. Engelhardt et al. (2017, p. 1275) argue that "individuals who process information more slowly suffer because alternative lexical argument structures and syntactic frames have substantially decayed once the disambiguating information is encountered".

The findings of the present study considering the relation between reading pace and response accuracy were somewhat mixed. In the first experiment, there were two regions (the conjunction and the ambiguous region) with lower RTs and two regions with higher RTs (the disambiguating word and the last word, i.e. subject NP) in the GP sentences where the comprehension questions were answered incorrectly. In the two other experiments, no effect of response accuracy on RTs was observed. Since the effects in the first experiment were rather weak (and two of these effects resulted from a post-hoc analysis) and they failed to replicate in Experiments 2 and 3, they are not included in the overall interpretation of the results. Taken together, the experiments suggest that the response accuracy is not (strongly) related to previous reading patterns in the examined GP sentences. It is possible that the difference between the findings here and findings in Engelhardt et al. (2017) or Stella & Engelhardt (2019) may be due to the different GP structures used. Both of these studies used object/subject GP sentences with either reflexive absolute transitive verbs (such as *While Anna dressed the baby that was small and cute spit up on the bed*) or optionally transitive verbs (such as *While Susan wrote the letter that was long and eloquent fell off the table*). We therefore tentatively propose that the ease of repair of the GP sentence may interact with the processing speed effects on response accuracy. That is, if a GP structure is very hard to process (and repair), processing speed effect is possibly diminished or neutralized.

It could be argued that the results of the present study are largely influenced by the experimental methods, namely the word-by-word ("moving window") self-paced reading paradigm. However, there are various reasons not to think that the results could be explained by this. First, the rate of correct answers on comprehension questions following the filler sentences was very high in all three experiments (i.e. 93.62%, 96.1%, 96.27%, respectively). Thus, the participants in these experiments were able to cope with the unusual way of reading quite well. Second, if indeed it was the method that influenced the results to a large extent, then such an influence should also be observed for the control sentences since these sentences were well-aligned with the GP sentences. However, the rate of incorrect answers for control sentences was generally low (with the proportion of incorrect responses typically being less than 10%). Third, the inability to reread could play a role in the processing of the sentences containing the syntactic ambiguity. However, Christianson et al. (2017) themselves claim that rereading is not related to the rate of incorrect answers on questions following GP sentences. Therefore, there is good reason to believe that the method used was not the cause of the response patterns found in these experiments.

The main finding of this study is that the resulting representations of certain GP sentences may be disrupted and confused. A range of difficulty levels related to the GP structures and resulting representations have been investigated. Some garden-paths are easy to process and the resulting representation is comprehended correctly. Others are harder and result in a lingering misanalysis. In the hardest cases, readers may often end up with a disrupted and incoherent representation of the sentence. However, not only GP sentences should be under scrutiny from this perspective, it might be the case that the processing system fails more often than typically admitted in the literature. This possibility should be addressed in future research, applying various types of constructions in various, typologically diverse languages.

Funding

This study was supported by the Charles University project PROGRES 4, Language in the Shiftings of Time, Space, and Culture.

References

Altmann, G. T., Garnham, A., & Dennis, Y. (1992). Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, *31*(5), 685–712. https://doi.org/10.1016/0749-596X(92)90035-V

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv Preprint ArXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.

Bever, T. G. (1970). The cognitive basis for linguistic structures. *Cognition and the Development of Language*, *279*(362), 1–61.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243.

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. https://doi.org/10.1080/17470218.2015.1134603

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407.

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

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology*, *70*(7), 1380–1405. https://doi.org/10.1080/17470218.2016.1186200

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Younger and Older Adults' "Good-Enough" Interpretations of Garden-Path Sentences. *Discourse Processes*, *42*(2), 205–238. https://doi.org/10.1207/s15326950dp4202_6

Clifton, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language*, *49*(3), 317–334. https://doi.org/10.1016/S0749-596X(03)00070-6

den Ouden, D.-B., Dickey, M. W., Anderson, C., & Christianson, K. (2016). Neural correlates of early-closure garden-path processing: Effects of prosody and plausibility. *Quarterly Journal of Experimental Psychology*, *69*(5), 926–949. https://doi.org/10.1080/17470218.2015.1028416

Engelhardt, P. E., Nigg, J. T., Carr, L. A., & Ferreira, F. (2008). Cognitive inhibition and working memory in attention-deficit/hyperactivity disorder. *Journal of Abnormal Psychology*, *117*(3), 591. https://doi.org/10.1037/a0012593

Engelhardt, P. E., Nigg, J. T., & Ferreira, F. (2017). Executive Function and Intelligence in the Resolution of Temporary Syntactic Ambiguity: An Individual Differences Investigation. *Quarterly Journal of Experimental Psychology*, *70*(7), 1263–1281. https://doi.org/10.1080/17470218.2016.1178785

Ferreira, F., Lau, E. F., & Bailey, K. G. (2004). Disfluencies, language comprehension, and tree adjoining grammars. *Cognitive Science*, *28*(5), 721–749. https://doi.org/10.1207/s15516709cog2805_5

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

Fodor, J. D., & Inoue, A. (1994). The diagnosis and cure of garden paths. *Journal of Psycholinguistic Research*, *23*(5), 407–434.

Fodor, J. D., & Inoue, A. (1998). Attach anyway. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 101–141). Springer.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325. https://doi.org/10.1016/0010-0277(78)90002-1

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jasinskaja, K., & Šimík, R. (in press). Slavonic free word order. In J. Fellerer & N. Bermel (Eds.), *The Oxford Guide to the Slavonic Languages*. Oxford University Press.

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040. https://doi.org/10.1080/17470218.2015.1053951

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H., & others. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037/0033-295X.101.4.676

Malyutina, S., & den Ouden, D.-B. (2016). What is it that lingers? Garden-path (mis)interpretations in younger and older adults. *Quarterly Journal of Experimental Psychology*, *69*(5), 880–906. https://doi.org/10.1080/17470218.2015.1045530

Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: Evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 280–285. https://doi.org/10.1037/a0014276

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 940–961. https://doi.org/10.1037/0278-7393.24.4.940

Pickering, M. J., Traxler, M. J., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language*, *43*(3), 447–475. https://doi.org/10.1006/jmla.2000.2708

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879. https://doi.org/10.1037/0021-9010.88.5.879

Pravdová, M., & Svobodová, I. (Eds.). (2014). *Akademická příručka českého jazyka*. Academia.

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. https://doi.org/10.1080/23273798.2017.1379606

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Siewierska, A., & Uhlířová, L. (1998). An overview of word order in Slavic languages. In A. Siewierska (Ed.), *Constituent order in the languages of Europe* (pp. 105–149). De Gruyter.

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, *69*(2), 104–120. https://doi.org/10.1016/j.jml.2013.04.001

Stella, M., & Engelhardt, P. E. (2019). Syntactic ambiguity resolution in dyslexia: An examination of cognitive factors underlying eye movement differences and comprehension failures. *Dyslexia*, *25*(2), 115–141. https://doi.org/10.1002/dys.1613

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, *105*(2), 477–488. https://doi.org/10.1016/j.cognition.2006.10.009

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316. https://doi.org/10.1016/S0749-596X(03)00081-0