Can computational models learn syntax? The learnability of the *wh*- and coordinate structure island constraints by artificial neural networks in Dutch

by

Michelle (M.J.P.F.) Suijkerbuijk

Research Master Linguistics and Communication Sciences

Department of Language and Communication, Radboud University

Nijmegen, July 21, 2022

Supervisors: dr. Stefan Frank and dr. Peter de Swart

Content

Summary

The ability to use language feels like the most natural thing in the world, but how is it possible? Human language learners are argued to need innate knowledge of a language, as research seems to show that they cannot learn language only from the language input they receive (Pearl, 2021). Interestingly, however, recent research has shown that artificial neural networks (ANNs), i.e. general learning systems without any knowledge of language built in, can acquire human-like grammatical knowledge solely based on the input they receive (Linzen & Baroni, 2021). One important problem with this recent research is that it is all performed in English (Mueller et al., 2020). Therefore, the current research project investigated whether an ANN can learn two syntactic island constraints, namely the *wh*- and coordinate structure island constraint, in Dutch in a way comparable to human native speakers.

First, it was established whether the *wh*- and coordinate structure island constraints exist in Dutch, and if so, to what extent human native speakers are sensitive to these constraints with an acceptability judgement task. Second, a Long Short-Term Memory (LSTM) network, trained on 12 million sentences extracted from the Dutch *Corpora Of the Web* (NLCOW14), assigned surprisal values to the same test sentences, indicating the extent to which a word was unexpected by the network, to assess whether its sensitivity to island violations was similar to that of the Dutch native speakers.

Unlike human native speakers, who demonstrate a clear sensitivity to *wh*- and coordinate structure island violations confirming the existence of these island constraints in Dutch, the LSTM network is not able to recognize *wh*- and coordinate structure islands and to block gap expectancies within these islands in Dutch. This suggests that input alone might not be enough to learn about syntactic island constraints, and that internal language knowledge or abilities might be necessary to learn about these constraints.

1. Introduction

For us as humans it is the most natural thing in the world that we are able to use language. A much debated question within the field of linguistics, however, is how this is actually possible; is this a capacity we are born with or do we learn it from all the language we hear in our environment? Sixty years ago, it was observed that language users seem able to go beyond what they can learn from the input they receive (Chomsky, 1965); they can create novel sentences and produce errors not present in the input (Bates, 2003). This is taken as evidence that the input data is insufficient or too impoverished on its own for children to make inferences possible about the correct use of language, commonly referred to as ‘Poverty of the Stimulus’ (henceforth PoS). It is impressive to see that language learners nonetheless learn to use language correctly. Some have argued that this is possible due to some other signal that helps speakers infer the correct use of language, besides the input they receive (Pearl, 2021). While some linguists argue for the use of prior linguistic knowledge or an innate language ability, the exact nature of this ‘other signal’ remains disputed. Therefore, the only claim largely agreed on is that language learners do not seem able to learn language just based on the input they receive (i.e. PoS). Interestingly, however, recent research has shown that artificial neural networks can learn grammar without any innate language knowledge.

Artificial neural networks (ANNs) are general learning systems, and when used as language models, they are exposed to large amounts of raw input text, which they encode to sequences of real numbers, i.e. vectors. As general learning systems, they have no built-in linguistic knowledge, and the vectors they process are not linguistic in nature either. All the more impressive that computational linguistics has recently shown that these ANNs can induce human-like grammatical knowledge (Linzen & Baroni, 2021). These findings breathe new life into the debate about how we as humans can learn language. That is because, to the extent that ANNs can learn syntax, an innate language ability is not required. Many studies have thus tried to find out to what extent ANNs are actually able to learn syntax. An important problem with this previous research is, however, that they almost never vary the input language these architectures receive; English dominates this field (Mueller et al., 2020). This is a problem because recent literature suggests that ANNs may have a performance advantage for English-like structural input caused by an overlap between English-like structures and a non-linguistic bias within the network (Davis & van Schijndel, 2020). The human language learning system must be universal, meaning that it needs to work equally well for all languages. If neural networks indeed work better for English, these learning systems would not be universal enough, falsifying the claim that ANNs can learn syntax without an innate language ability.

Therefore, the current research project investigated whether ANNs can learn syntactic constraints in Dutch in a way comparable to human Dutch speakers. Specifically, I examined the learnability of syntactic island constraints by neural networks in Dutch, and compared the network’s performance to that of experimentally tested native Dutch speakers. In Section 2 and 3, I will provide a theoretical background on the current research topic. Specifically, in Section 2, I will elaborate more on the theory and experimental research published about syntactic island constraints, and in Section 3, I will explain ANNs in more detail and discuss previous research performed with these ANNs on syntactic island constraints. In Section 4, I will introduce the current research project, which will be discussed in detail in the sections following that. The acceptability judgement task performed by native speakers will be reported on and its results will be discussed in Section 5, and the experiment with the ANN will be described and discussed in Section 6. The native speakers and neural network will be compared in Section 7, in which the strengths, limitations and implications of this research project will also be discussed.

2. Syntactic island constraints

Many of the languages in the world exhibit dependency relations between two elements in a sentence. To form such a dependency in English and Dutch, one element (i.e. the filler) dislocates from one position to another in the sentence leaving behind a gap. This is illustrated in (1).[1]

(1) a. Mary saw Bill.
b. Who$_i$ did Mary see ___$_i$?

(Abeillé et al., 2020, p. 1)

Generally, it is assumed that these filler-gap dependencies are not constrained by the distance between the filler and its gap in *wh*-constructions; the filler and the gap can be separated by any number of words and clause boundaries (Sprouse & Hornstein, 2013; Abeillé et al., 2020). They do, however, seem constrained by the type of structure the filler is moved out of. Previous research has shown that the examples of filler-gap dependencies in (2) are perceived as unacceptable by most native English speakers (Hofmeister & Sag, 2010).[2] These structures, therefore seem gap-resistant (Sprouse et al., 2012; Sprouse & Hornstein, 2013).

(2) *Wh*-island
a. *What$_i$ do you wonder [$_{wh}$ whether John bought ___$_i$]?
Complex Noun Phrase island (henceforth: CNP island)
b. *What$_i$ did you make [$_{NP}$ the claim that John bought ___$_i$]?
Subject island
c. *What$_i$ do you think [$_{\text{subject NP}}$ the speech about ___$_i$] interrupted the TV show?
Sentential subject island
d. *What$_i$ did [$_{\text{subject CP}}$ that John wrote ___$_i$] offend the author?
Adjunct island
e. *What$_i$ do you worry [$_{\text{adjunct}}$ if John buys ___$_i$]?
Relative clause island
f. *What$_i$ did you meet the scientist [$_{\text{relative clause}}$ who invented ___$_i$]?
Coordinate structure island
g. *What$_i$ did John buy [$_{\text{coordination}}$ a shirt and ___$_i$]?
Left-branch island
h. *Which$_i$ did John borrow [$_{NP}$ ___$_i$ book]?

(Sprouse et al., 2012, p. 83)

[1] The gap is represented by underscores and the *wh*-filler and the gap are co-indexed with an 'i'.
[2] Ungrammaticality or unacceptability is marked by an asterisk.

Ross (1967) was the first one to give these gap-resistant structures a name: syntactic islands. Newmeyer (2016) defined these as “a syntactic domain containing an element that cannot be coindexed with an element outside of that domain” (p. 188), which simply means that it is ungrammatical to create a dependency between an element in the sentence and an element in the syntactic island configuration. In the existing literature, this ungrammaticality is also called the ‘island effect’. While the island effect can occur in different types of structures, all illustrated in (2), the discussion in the current thesis will focus only on *wh*- and coordinate structure islands. This choice will be motivated in Section 3.3.

2.1 The source of syntactic island effects

A much-debated question within the literature on syntactic island effects regards their source; why do island effects arise? Analyses generally fall into three categories, namely syntactic, extra-syntactic competence-based, and extra-syntactic performance-based approaches (Newmeyer, 2016).

In general, linguists agree that syntactic island constraints are unlikely to be observed in the input that children receive. Therefore, after their first discovery, syntacticians tried to account for their existence by constructing different innate grammatical constraints (i.e. island constraints) (Hofmeister & Sag, 2010; Newmeyer, 2016). Chomsky (1964) was one of the firsts to create a general constraint on long-distance dependencies, namely the A-over-A condition. According to this condition, an element of a certain category cannot be extracted out of a phrase of the same category. Chomsky used this condition to explain why for example (3b) must be ungrammatical; an NP cannot be extracted out of another NP (Hofmeister & Sag, 2010).

(3) a. John kept the car in the garage.
b. *[$_{\text{NP}}$ What$_{\text{i}}$] did John keep [$_{\text{NP}}$ the car in ___$_{\text{i}}$]?
(Newmeyer, 2016, p.190)

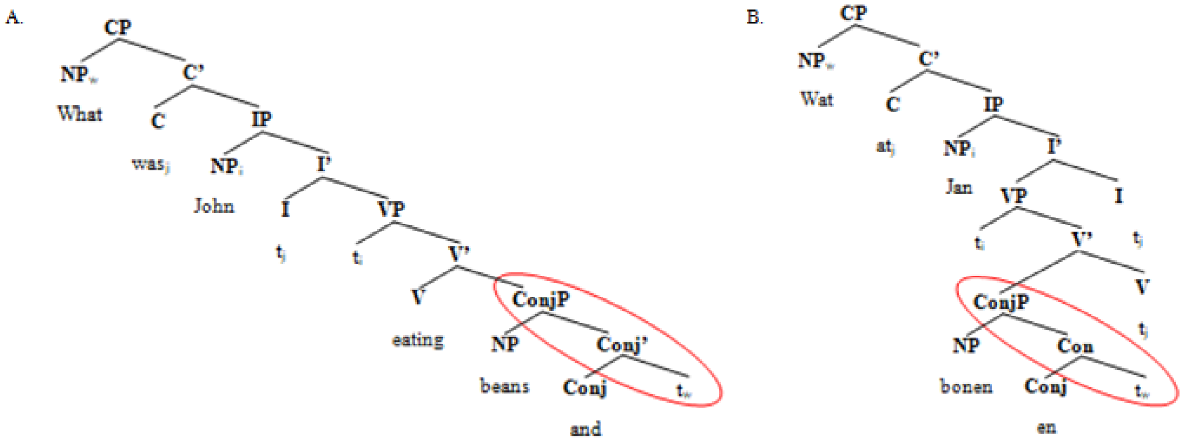
However, Ross (1967) noted that the A-over-A condition predicted grammaticality in sentences that are perceived as unacceptable, such as the coordinate structure island in (4).

(4) a. You think Sandy photographed the castle and Chris visited the dignitaries.
b. *[$_{\text{NP}}$ Which dignitaries$_{\text{i}}$] do you think [$_{\text{Coordinate structure}}$ [$_{\text{IP}}$ Sandy photographed the castle] and [$_{\text{IP}}$ Chris visited ___$_{\text{i}}$]]?
(Hofmeister & Sag, 2010, p. 5)

The NP ‘which dignitaries’ is moved out of the IP ‘Chris visited’, which should be possible under the A-over-A condition as these elements are from a different category. Yet, it does not seem possible to move (part of) a full conjunct out of coordination in any language (Liu et al., 2022). As a reaction, therefore, Ross (1967) created universal, and thus language-independent, constraints for complex noun phrases, left branches, sentential subjects and coordinate structures, which restrict movement out of these specific types of structures. The ‘Coordinate Structure Constraint’, for example, entails that “no conjunct in a coordinate structure may be moved" nor "any element in a conjunct" (Newmeyer, 2016:191). With this constraint, it is clear why the coordinate structure island configurations in Figure 1A for English and in Figure 1B for Dutch are perceived as highly ungrammatical. In these two examples, in both languages, an NP is moved out of a coordinate structure, represented by the Conjunction Phrase (ConjP).

Figure 1

Syntactic tree structure of the English sentence 'what was John eating beans and?' in A and of the Dutch equivalent 'wat at John bonen en?' in B. The conjunct out of which an element is moved is circled in red.



For *wh*-islands, Ross (1967) decided not to create a distinct constraint. He believed that any constraint was and would be too strong for this island type as it has many grammatical exceptions. For example, Chomsky's (1964) A-over-A condition would rule out the sentences in (5), while Ross found these sentences totally acceptable. Therefore, he argued that much more work needed to be done to create a weaker constraint to explain *wh*-island effects.

(5) He told me about a book$_i$ which I can't figure out [$_{\text{wh-phrase}}$ whether to buy ____$_i$ or not/how to read ____$_i$/where to obtain ____$_i$/what to do about ____$_i$].

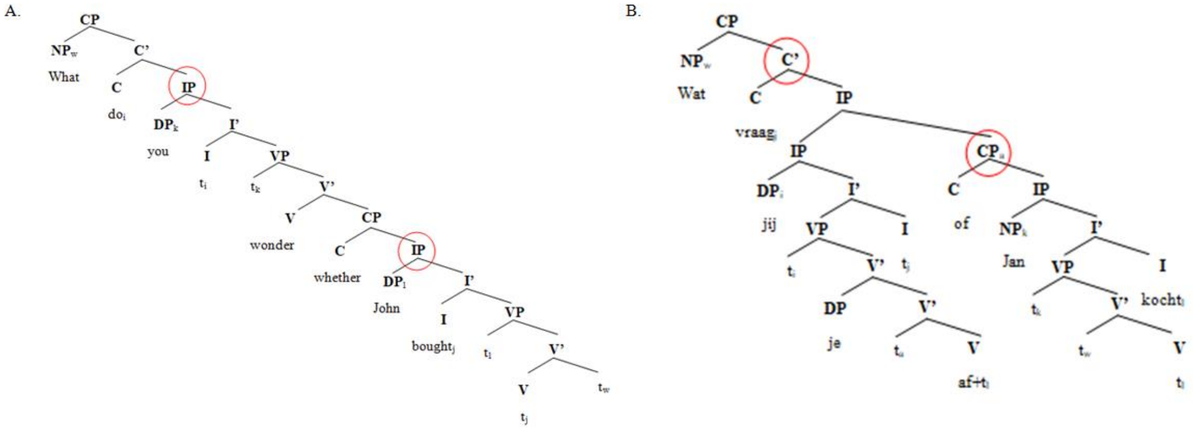
(Ross, 1967, p. 19)

However, exceptions were not only found for *wh*-islands, but almost for every island type. The only constraint still widely agreed upon is the Coordinate Structure Constraint (Liu et al., 2022). Chomsky (1973) tried to account for the exceptions by combining all existing constraints at that time into a general principle of Subjacency. This principle can not only account for *wh*-islands, but also for CNP, subject and left-branch islands.

Subjacency prohibits movement that crosses more than one so-called bounding node or barrier, represented by the IP and the DP/NP in English (Hofmeister & Sag, 2010; Newmeyer, 2016). Figure 2A shows how Subjacency can account for the *wh*-island effect in English.

Figure 2

Syntactic tree structure of the English sentence 'what do you wonder whether John bought?' in A and of the Dutch equivalent 'wat vraag jij je af of Jan kocht?' in B. The two bounding nodes crossed are circled in red.



In Figure 2A, the object of 'bought' (i.e. the DP 'what') moves from the complement of the embedded VP to the specifier of the main CP, crossing two bounding nodes, namely the embedded and the main IP. Hence, this sentence is perceived as ungrammatical.

To explain any cross-linguistic differences, the bounding nodes or barriers were later parametrized, so that each language could set the bounding nodes differently (Rizzi, 1982). For Dutch, for instance, the DP/NP and the CP would be better as bounding nodes. That is because the CP serves as the clause boundary in Dutch, while the IP fulfills this role in English. In Figure 2B, it can be seen that the object of *kocht* 'bought' (i.e. the DP *wat* 'what') moved from the complement of the extraposed embedded VP to the specifier of the main CP, crossing two bounding nodes, namely the extraposed CP and the main C'. Therefore, according to Subjacency, this Dutch sentence should be perceived as ungrammatical.

While the principle of Subjacency thus seems able to account for the observed *wh*-island effect in different languages, again other linguists found exceptions to the rule. The example in (6), for example, illustrates that it can be acceptable to cross more than one bounding node.

(6) Which reports$_i$ does [$_{IP}$ the government prescribe [$_{NP}$ the height of [$_{NP}$ the lettering on ___$_i$]]].

(Hofmeister & Sag, 2010, p. 7)

Therefore, linguists started to supplement the existing syntactic constraints with extra-syntactic factors, such as semantic, discourse or processing components. According to Newmeyer (2016), the reason for this was threefold: (1) the data of syntactic islands was much more complex than thought at first; (2) the exceptions that needed to be made for the constraints to work made the theories not as minimal as they should be according to the Minimalist Program (Chomsky, 1993); and (3) many linguists came to believe that pure syntactic constraints could not solely explain all the existing data. Consequently, the extra-syntactic approach to syntactic island effects came into existence.

According to this extra-syntactic approach, island-violating sentences are in fact grammatical structures. They are only perceived as unacceptable because of non-syntactic factors, such as discourse and processing components, creating the so-called competence-based and performance-based approaches.

First, the competence-based approach claims that island effects arise due to the information-structure properties of syntactic island configurations and/or of the filler. Specifically, it argues that island effects arise when the information-structure properties of the island and the filler clash (Newmeyer, 2016). Numerous discourse principles have tried to capture the syntactic island data, focusing on different discourse elements (e.g. dominance, topichood, salience, and relevance) (for a detailed discussion of these approaches see Liu et al., 2022). Explaining every one of these approaches is beyond the scope of the current thesis, so only one example will be discussed here. One of the most recent and compelling principles was created by Goldberg (2014): "Backgrounded Constructions are Islands". Movement makes an element prominent in the discourse. Consequently, to avoid a clash of information-structure properties of two elements, movement can only take place out of another prominent element. It is thus never possible to move out of a backgrounded constituent, such as subjects, relative clauses and definite nouns. Therefore, backgrounded constituents are gap-resistant and, according to Goldberg's principle, can be labelled as islands (Abeillé et al., 2020; Newmeyer, 2016).

Second, the performance-based approach claims that some parts of the parsing system, also involved in regular sentence production and comprehension, can affect the acceptability of syntactic island configurations (e.g. encoding aspects of a syntactic structure in and retrieving it from our working-memory) (Liu et al., 2022; Sprouse & Hornstein, 2013). This approach was mainly developed to explain the graded acceptability that was found for certain island constructions; not all constructions in (2) are rated as completely unacceptable, but some more than others (Abeillé et al., 2020; Hofmeister & Sag, 2010). The purely syntactic constraints have trouble explaining this graded acceptability as most assume that the grammar is binary; something is either grammatical or ungrammatical, but cannot be in-between. The performance-based approach tries to explain this graded acceptability by involving certain processing factors. Some structural features, such as a bare *wh*-filler or the syntactic distance between filler and gap, can make an island configuration into a complex structure. Keeping a bare filler in our working-memory, or keeping any filler in our working-memory when simultaneously processing many other words, can cause a processing overload, which can in turn result in perceived unacceptability (Abeillé et al., 2020; Hofmeister & Sag, 2010; Liu et al., 2022; Newmeyer, 2016). Therefore, according to the performance-based approach, syntactic island configurations are not ungrammatical, they are simply too hard to process.

While these three approaches thus all try to explain syntactic island effects in a different way, it is still an on-going discussion which of these approaches can explain the source of island effects the best. This discussion ties in nicely with the previously mentioned on-going debate about the assumed innate language ability. Do we possess innate syntactic knowledge and thus (possibly) also innate syntactic island constraints, or do we learn syntax from the input and could island constraints therefore either be explained syntactically or extra-syntactically? Consequently, the current research about the learnability of syntactic island constraints by ANNs in Dutch can make a relevant contribution to both of these debates.

2.2 Experimental investigations of syntactic island constraints

Numerous experimental investigations have been performed into syntactic island effects in different languages (e.g. Hofmeister & Sag, 2010; Pham et al., 2020; Sprouse et al., 2012 (for English); Pañeda et al., 2020 (for Spanish); Keshev & Meltzer-Assher, 2018 (for Hebrew); Kush et al., 2019 (for Norwegian)). These studies show that most languages exhibit island effects, but that these effects differ between and within languages and island types. For example, because the unacceptability of an island violation can differ between island types, island types are generally divided into 2 classes: strong and weak islands (Newmeyer, 2016). In general, violations of strong island constraints are always perceived as highly unacceptable, while the degree of unacceptability of weak islands can vary based on several semantic and processing factors (e.g. discourse status or the syntactic complexity of the *wh*-filler).

An example of a strong island is the coordinate structure island. Ross (1967) already stated that his Coordinate Structure Constraint is universal and thus applies to all languages. Fifty-five years later, not a single linguist seems to disagree with him: “it does not seem possible to extract one or more full conjuncts, in any language” (Liu et al., 2022, p. 503). This also means that coordinate structure islands are (almost) never included in experimental investigations of island effects. Consequently, the current research will be the first to experimentally investigate the Coordinate Structure Constraint with human native speakers.

Wh-islands, on the other hand, are a clear example of weak islands in English. Generally, *wh*-islands are perceived as unacceptable, but can be ameliorated by using a complex *wh*-filler instead of a bare one (i.e. a lexical phrase) as in (7c) (Hofmeister & Sag, 2010; Sprouse et al., 2016) and/or by adding a discourse context as in (7) (Kush et al., 2019).

(7) Context: Albert learned that the managers dismissed the employee with poor sales after the annual performance review.

a. Who$_i$ did Albert learn [$_{CP}$ that they dismissed ___$_i$ after the annual performance review]?

b. *Who$_i$ did Albert learn [$_{wh}$ whether they dismissed ___$_i$ after the annual performance review]?

c. $^?$Which employee$_i$ did Albert learn [$_{wh}$ whether they dismissed ___$_i$ after the annual performance review]?

(Hofmeister & Sag, 2010, p. 44)

Therefore, in contrast to coordinate structure islands, there have been numerous experimental investigations into *wh*-island effects covering different languages. Dutch, however, remains underrepresented in this research and in research on syntactic island effects in general. Consequently, not much is known about whether syntactic island constraints exist in Dutch, and if so, to what extent Dutch speakers are sensitive to them. Beljon et al. (2021) is one of the few, if not the only, study that empirically investigated whether the *wh*-island constraint exists in Dutch and, if so, to what extent speakers are sensitive to it. Specifically, they investigated the acceptability of *wh*-islands in Dutch and the effect of (1) the complexity of the filler phrase (‘who’ vs. ‘which girl’) and (2) adding a preceding discourse context. Their results showed that the *wh*-island constraint exists in Dutch as speakers showed a sensitivity to this constraint; *wh*-islands were rated less acceptable on a 7-point scale than sentences without islands. Moreover, while neither filler complexity nor discourse context had an effect in isolation, their combination did constitute a significant effect; *wh*-islands with a complex filler and a preceding discourse context were rated as more acceptable than islands without, but this difference in acceptability was only minimal.

In addition to the published research by Beljon et al. (2021), I collected unpublished data on syntactic islands during the course 'Syntax in the Lab' in June 2021 (Suijkerbuijk, 2021). Specifically, the acceptability of *wh*-islands and whether making the filler more complex affected this acceptability were investigated in Dutch. Similar to Beljon et al. (2021), these results showed that the *wh*-island constraint exists in Dutch as its speakers show a strong sensitivity to *wh*-island violations.

With the current research project, the necessary experimental data on *wh*- and, more importantly, coordinate structure islands will be gathered with an acceptability judgement task to (1) provide relevant data to the experimental research on syntactic island effects in Dutch, and (2) enable a direct comparison between human Dutch speakers and the artificial neural network.

3. Computational linguistics and syntactic island constraints

In the past decade, artificial neural networks have commonly been used for tasks (e.g. machine translation or reading comprehension) within the research area of Natural Language Processing (NLP), which investigates how computational models can understand and produce natural language (Chopra et al., 2013; Linzen & Baroni, 2021). This is a remarkable fact for many linguists, because these networks don't possess the traits considered necessary for language acquisition, such as built-in linguistic knowledge. Still, recent research has shown that ANNs are able to accurately learn about, for example, number agreement (i.a. Goldberg, 2019; Gulordava, 2018), and garden paths (Frank, 2021; Frank & Hoeks, 2019; Futrell, 2019). Not all syntactic phenomena can successfully be learned yet, however. Syntactic island constraints, for instance, still receive mixed results in English. *Wh*-, CNP, coordinate structure, adjunct and left branch islands are, for example, successfully learned in most studies, but negative phrase, relative clause and (sentential) subject islands only partially or not at all (Chaves, 2020; Chowdhury & Zamparelli, 2018; Wilcox et al., 2018; Wilcox et al., 2019b; Wilcox et al., 2021).

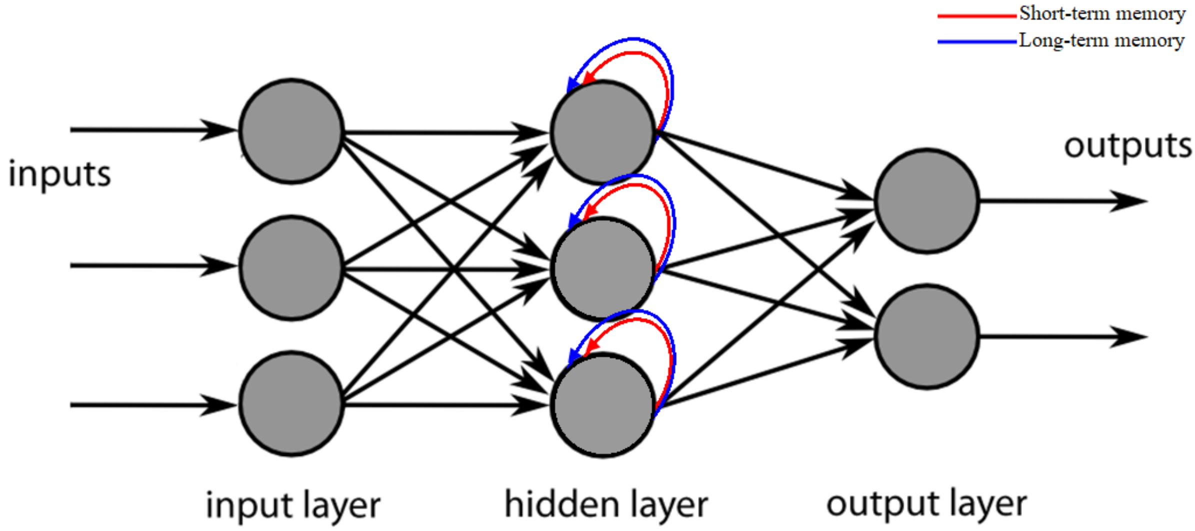
In this section, I will first explain more about the internal structure of artificial neural networks, and elaborate on possible concerns about using these networks for research on language acquisition. Last, I will go in more detail about previous research using ANNs to learn syntactic island constraints.

3.1 Artificial neural networks

Our brain possesses about 86 billion neurons, which are connected to each other and share information to learn (Kemmerer, 2015). Artificial neural networks are inspired on the architecture of our brain and its neural network, and aim to simulate this learning process (Walczak & Cerpa, 2003). Simply put, the ANN maps input from (hidden) layer to (hidden) layer to eventually compute an output based on the information it gathered in these hidden layers. These are also called feedforward networks, of which the architecture is illustrated in Figure 3.

Figure 3

Simplified architecture of an artificial neural network, with the red arrows representing the feedback loop of a simple recurrent neural network and the blue errors representing the long-term memory present in the LSTM network.



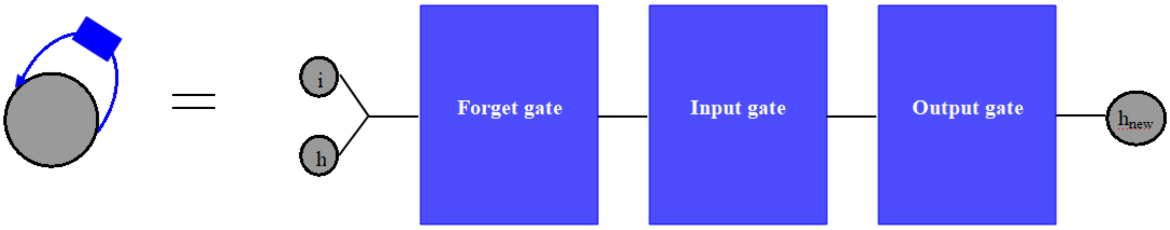
Note. Red and blue arrows added myself. From “A Simple Starter Guide to Build a Neural Network,” by J. Hu, 2018, January 20, *Towards Data Science*. (https://towardsdatascience.com/a-simple-starter-guide-to-build-a-neural-network-3c2cf07b8d7c).

Feedforward networks are, however, only able to encode individual words and not sequences of words, which is why Recurrent Neural Networks (RNNs) were introduced (Elman, 1990). Contrary to these feedforward networks, RNNs possess a memory that contains information from its own former internal state. When processing a sentence, it can use this information to compute outputs. This memory is captured in the so-called feedback loop, illustrated by the red arrows in Figure 3, and can help the RNN to learn the relations between words in a sentence (Linzen & Baroni, 2021; Saeed, 2021). The problem with the memory of RNNs is that it can only be used to learn short dependencies in sentences as it only contains the former hidden state, i.e. the information from the hidden layers processed right before. Although this hidden state is dependent on all formerly processed hidden states due to the feedback loop, this dependency is only indirect and thus difficult to use when learning dependencies stretched out over a longer linear distance (Wilcox et al., 2021). Therefore, Long Short-Term Memory (LSTM) models were designed.

As its name already suggests, LSTM models possess both a long-term and a short-term memory component. These memory components are represented by the blue and red arrows respectively in Figure 3. Moreover, it uses so-called gates. The architecture of the long-term memory with these gates is illustrated in Figure 4. Both the memory components and the associated gates are used to process a sentence.

Figure 4

The architecture of the long-term memory component of an LSTM network, illustrating the forget, input and output gate. 'i' stands for the current input, 'h' for the previous hidden state, and 'h_{new}' for the newly created hidden state.



When the input comes in, it is combined with the previous hidden state and fed to the forget gate. In this gate, it is determined what information is important to retain and what can be forgotten. Next the current input and previous hidden state go through the input gate, in which it is decided which information is most important to update. Last, when send to the output gate, a new hidden state is created based on the information gathered in the forget and input gate (Phi, 2018). The development of this Long-Term Memory component allows the LSTM model to have more information at its disposal and the introduction of the three gates ensures more control over this information. Consequently, the LSTM is able to represent long-distance dependencies.

In general, within the field of computational psycholinguistics, the types of ANNs described above are used in one of three possible settings. First, it can be used as a classifier, which gives a discrete label as output for an input sequence, such as 'acceptable' or 'unacceptable'. A classifier is trained on a set of sentences already annotated for their acceptability, making this setting supervised. Second, there is the sequence-to-sequence setting, in which the network produces an output sequence in response to an input sequence. This setting is commonly used in machine translation. Third, neural networks can also be used as language models. A language model outputs a probability value for each input word, based on its preceding context. In this setting, the network is trained on a set of sentences without any information annotated, making it unsupervised (Linzen & Baroni, 2021).

To investigate whether an artificial neural network is able to learn the *wh*-island and coordinate structure island constraints, it must be able to track long-distance dependencies. Therefore, a Long Short-Term Memory model was used in the current research project. To see how the network performs in specific regions of the sentence and without any supervision, the LSTM network was used in the language model setting.

3.2 Computational investigations of syntactic island constraints

3.2.1 Syntactic or extra-syntactic factors at play?

One of the first computational investigations on the learnability of long distance dependencies concerned the agreement between a subject and a verb (Linzen et al., 2016; Gulordava et al., 2018). These successful investigations showed that, when RNNs are presented by the sequence 'The key to the cabinets…', they assign a higher probability to the correct singular verb form 'is' than to the incorrect plural verb form 'are'. To strengthen the claim that the dependency between a subject and a verb can be maintained by RNNs, other studies also showed that the network was also able to choose the correct verb form in languages other than English and in semantically implausible sentences. Subject-verb agreement is, however, a syntactic phenomenon that frequently occurs in the set of sentences the network is trained on. This makes it easy to claim that this phenomenon can be learned from the input only, without any innate syntactic constraints necessary. To strengthen the claim that RNNs can acquire different long distance dependencies without any innate syntactic knowledge, so based on the input only, it is also important to investigate dependencies not often seen in the training data set. If these dependencies cannot be learned by the RNN, it shows that some innate syntactic knowledge is necessary to learn about these long distance dependencies. On the other hand, if the RNN is able to learn these dependencies, it demonstrates that the input is enough, even if the phenomenon itself does not often occur in this input. Therefore, investigations followed in which syntactic phenomena were researched involving long-distance dependencies not often seen in the training data set (Chowdhury & Zamparelli, 2018). One of these phenomena, central to the current research project, is the constraints on filler-gap dependencies, namely syntactic island constraints. In this section, I will discuss the previously performed computational investigations on the learnability of these constraints by LSTM networks, specifically trained on a language modelling objective.

Chowdhury and Zamparelli (2018) were one of the firsts to investigate the learnability of syntactic island constraints by LSTM networks. Specifically, they examined whether these networks could detect ungrammatical sentences based on their total-sentence probability. In their examination, they looked at regular *wh*-extractions and subject and relative clause island violations. First, they started by looking at regular *wh*-extractions, without any island configurations, such as the minimal pair in (8).

(8) a. Which candidate$_i$ should the students discuss ___$_i$?
b. *Which candidate$_i$ should the students discuss {him / something else / this candidate}$_i$?

(Chowdhury & Zamparelli, 2018, p. 137)

The results showed that LSTMs can recognize ungrammatical *wh*-extractions, such as the one in (8b), but that this became more difficult when the sentence contained more embedding layers (e.g. ‘which candidate$_i$ did the teacher think the students should discuss __$_i$?’). Now that it was clear that LSTMs could generally recognize regular *wh*-extractions, Chowdhury and Zamparelli (2018) investigated whether LSTMs could also represent the constraints on these extractions, namely the subject and relative clause island constraints in (9).

(9) Subject island
a. $^?$Who$_i$ did John see [$_{object}$ a classmate of ___$_i$]?
b. *Who$_i$ did [$_{subject}$ a classmate of ___$_i$] ruin John?
Relative clause island
c. *Which girl$_i$ did John see [$_{relative\ clause\ in\ object}$ the person that dated ___$_i$]?
d. *Which girl$_i$ did [$_{relative\ clause\ in\ subject}$ the person that dated ___$_i$] see John?

(Chowdhury & Zamparelli, 2018, p. 138)

The results showed that LSTMs seemed to have learned to recognize gap-resistant island configurations; both subject and relative clause islands were assigned lower total-sentence probability than sentences with regular object extraction. Interestingly, however, the networks assigned a similar total-sentence probability to yes/no-questions, such as the one in (10), which does not have *wh*-movement and is thus never able to exhibit island effects.

(10) Did John see the person that dated Mary?
(Chowdhury & Zamparelli, 2018, p. 141)

Therefore, Chowdhury and Zamparelli (2018) suggest that the unacceptability of subject and relative clause islands cannot be due to the simple fact that they contain a gap-resistant island, but that it is the result of the cumulative effect of a syntactically complex structure and the position of that structure. Subject islands, such as the subject 'a classmate of Tim', are more complex than for instance the subjects 'Tim, a classmate' or 'Tim's classmate', and relative clauses are argued to be structurally complex as well. Chowdhury and Zamparelli, furthermore, suppose that these complex structures lead to ambiguity, which makes the networks more uncertain, explaining the low probabilities assigned to these structures. Moreover, the position of such a complex structure also affects the probability; it is better to have a complex structure at the end of the sentence than at the start of it. At the start of the sentence, a complex and potentially ambiguous structure can confuse the network, and this confusion will influence how the network processes the rest of the sentence, leading to a lower total-sentence probability.

In conclusion, Chowdhury and Zamparelli (2018) thus suggest that neural networks do not use their 'sense' of grammaticality when processing sentences with ungrammatical *wh*-extraction and syntactic islands, but that processing factors are at play, such as the number of embedding layers of the sentence, its syntactic complexity and the position of this complexity.

After this investigation, Wilcox et al. (2018) followed with a paper in which they come to an entirely different conclusion. Before discussing their results, however, it is first important to discuss the experimental design they used. This was developed to investigate whether RNNs can learn regular filler-gap dependencies and island constraints, specifically by zooming in on the learnability of different predictions assumed to be made by the grammar. This design was later also used in various other papers, that will be discussed in this section as well.

Wilcox et al.'s (2018) experimental design constitutes a 2×2 interaction design, based on two predictions assumed to be made by the grammar: (1) gaps require fillers, and (2) fillers require gaps. Consequently, PRESENCE OF GAP and PRESENCE OF FILLER are crossed, and the resulting four conditions can be found in Table 1.

Table 1

Regular filler-gap dependency in the four conditions of Wilcox et al.'s (2018) interaction design.

Item	Gap?	Filler?	Example sentence
a	Yes	Yes	I know what$_i$ he said that the lion devoured ___$_i$ at sunrise.
b	Yes	No	*I know that he said that the lion devoured ___$_i$ at sunrise.
c	No	Yes	*I know what$_i$ he said that the lion devoured a gazelle$_i$ at sunrise.
d	No	No	I know that he said that the lion devoured a gazelle at sunrise.

If the network indeed assumes that gaps require fillers, gaps should be more surprising when no *wh*-filler is present. This means that 'at sunrise' should be more surprising in item (b) than in item (a). This effect is also called the *wh*-effect in the +Gap condition (Wilcox et al., 2021). Furthermore, if the network also assumes that fillers require gaps, filled argument positions should be more surprising when a *wh*-filler is present. This means that 'a gazelle' should be less surprising in item (d) than in item (c). In humans, this effect is called the filled-gap effect, but for neural networks it is referred to as the *wh*-effect in the –Gap condition (Wilcox et al., 2021). How well both assumptions are learned can be measured with the difference of differences, i.e. the full licensing interaction ((1b)−(1a))−((1d)−(1c)) (Wilcox et al., 2018; Wilcox et al., 2019b).

To investigate whether neural networks are also able to learn about island constraints, STRUCTURE (non-island vs. island) is added to this interaction design. The four island conditions can be found in Table 2.

Table 2

Wh-islands in the four conditions of Wilcox et al.'s (2018) interaction design.

Item	Gap?	Filler?	Example sentence
e	Yes	Yes	*I know what$_i$ he said whether the lion devoured ___$_i$ at sunrise.
f	Yes	No	*I know that he said whether the lion devoured ___$_i$ at sunrise.
g	No	Yes	*I know what$_i$ he said whether the lion devoured a gazelle$_i$ at sunrise.
h	No	No	I know that he said whether the lion devoured a gazelle at sunrise.

Unlike in regular filler-gap dependencies, when an island configuration contains a gap, the presence of a *wh*-filler should not affect the network's expectations; the network should never expect a gap inside an island, as this is ungrammatical. The *wh*-effect in the +Gap condition should thus be close to zero. Wilcox et al. (2018) argue in similar vein for island configurations with filled argument positions and no gaps; the presence of a filler should not affect the network's expectations as the neural network will always expect a filled argument position inside an island. According to Wilcox et al. (2018), the *wh*-effect in the –Gap condition should thus also be close to zero. While I agree that the network should always expect a filled argument position and never a gap inside an island, I think the network will still be affected by the presence of a *wh*-filler in sentences without a gap inside an island. While the network should never expect a gap inside an island, coming across a *wh*-filler at the start of the sentence should give rise to the expectation of a gap somewhere else. When encountering the period whilst not having encountered a licit gap, there should be a spike in surprisal, as these sentences are completely ungrammatical. This would mean that the *wh*-

effect in the –Gap condition should be negative. Regardless of the interpretation of the *wh*-effect in the –Gap condition, however, the learnability of island constraints can still be assessed by comparing the full licensing interaction of non-islands to that of island violations. If island constraints are learned correctly, a significant decrease in the licensing interaction is expected (Wilcox et al., 2018; Wilcox et al., 2019b).

Contrary to Chowdhury & Zamparelli (2018), Wilcox et al. (2018) showed with this interaction design that different LSTM networks could successfully learn *wh*-, adjunct and CNP islands. Subject islands, however, proved to be too difficult to learn correctly. Wilcox et al. (2018) argue that the difference between their and Chowdhury and Zamparelli's investigations can be attributed to the experimental design used in both studies. First of all, Chowdhury and Zamparelli used sentence schemata to generate numerous sentences with exactly the same structure but different content words, while Wilcox et al. (2018) carefully made the test items themselves paying attention as to whether each item tested the correct phenomenon. Moreover, Wilcox et al.'s (2018) non-island and island sentences were arguably equally complex, in contrast to those used by Chowdhury and Zamparelli. All the reasons above could have contributed to the different results obtained by both studies. Later, however, Wilcox et al. (2019b) found a way to directly test whether there are syntactic or extra-syntactic factors at play.

3.2.2 A control study

While Wilcox et al. (2018) argue that LSTM networks can learn the syntactic *wh*-, adjunct and CNP island constraints, Chowdhury and Zamparelli (2018) suggest that these networks are affected by processing factors, namely the syntactic complexity of islands and the position of this complex structure. Wilcox et al. (2019b) designed a control study to test both explanations. As Chowdhury and Zamparelli (2018) argue that neural networks are simply not able to thread information through syntactically complex structures (i.e. islands), Wilcox et al. (2019b) included sentences in which expectations for gendered pronouns needed to be established and maintained through complex syntactic structures. In these sentences, nouns with an unambiguous gender bias (e.g. 'actress') created the expectation that a pronoun further on in the sentence would match the noun's gender (i.e. 'her'). Consequently, GENDER MISMATCH (match vs. mismatch) and ISLAND (non-island vs. island) were crossed to create test items, of which examples can be found in Table 3.

Table 3

Example sentences of control items used in Wilcox et al. (2019b). The gendered noun and pronoun are boldfaced.

Gender mismatch?	Island?	Example sentence
No	No	The **actress** said that they insulted **her** friends.
Yes	No	The **actress** said that they insulted **his** friends.
No	Yes	The **actress** said whether they insulted **her** friends.
Yes	Yes	The **actress** said whether they insulted **his** friends.

A gendered expectation effect was calculated by taking the surprisal difference between the levels of GENDER MISMATCH, e.g. the difference between item (a) and (b). Wilcox et al. (2019b) showed that this gendered expectation effect did not reduce within an island configuration as compared to non-islands. This means that the expectation for gender can be threaded through the syntactically complex island configuration, and that it can be assumed that other information can be threaded through island as well.

After establishing that any results found were not caused by processing factors, Wilcox et al. (2019b) showed that *wh*-, adjunct, CNP and left branch islands could be learned successfully by two LSTM networks. While the networks also successfully learned not to expect a gap within coordinate structure islands, they still kept some expectation for a gap, completely unlike human behaviour (Liu et al., 2021). Moreover, similar to the LSTM models in Wilcox et al. (2018), the networks still did not show any sensitivity to (sentential) subject islands. While these results generally show that LSTM networks are able to learn syntactic constraints on filler-gap dependencies instead of simply being sensitive to their complexity, they also suggest that the networks are not completely human-like and that they are not able to learn all constraints successfully yet.

Similar to Wilcox et al. (2019b), Wilcox et al. (2019a) also designed a control condition for the complexity explanation. They examined whether neural networks are also able to license a gap over a syntactic island construction. An example of an item set with such a control item can be found in Table 4 and Table 5.

Table 4

Example sentences of CNP islands used in Wilcox et al. (2019a). Only item (a) is taken directly from Wilcox et al. (2019a). The other items are self-made adaptions.

Item	Gap?	Filler?	Example sentence
a	Yes	Yes	*I know who$_i$ the count that insulted ___$_i$ on the balcony talked with the hostess.
b	Yes	No	*I know that the count that insulted ___$_i$ on the balcony talked with the hostess.
c	No	Yes	*I know who$_i$ the count that insulted the woman$_i$ on the balcony talked with the hostess.
d	No	No	I know that the count that insulted the woman on the balcony talked with the hostess.

Table 5

Example sentences of the control condition used in Wilcox et al. (2019a). Only item (a) is taken directly from Wilcox et al. (2019a). The other items are self-made adaptions.

Item	Gap?	Filler?	Example sentence
a	Yes	Yes	I know who$_i$ the count that insulted the hostess talked loudly with ___$_i$ on the balcony.
b	Yes	No	*I know that the count that insulted the hostess talked loudly with ___$_i$ on the balcony.
c	No	Yes	*I know who$_i$ the count that insulted the hostess talked loudly with the woman$_i$ on the balcony.
d	No	No	I know that the count that insulted the hostess talked loudly with the woman on the balcony.

Wilcox et al. (2019a) tested the learnability of adjunct and CNP islands using the interaction design introduced in Wilcox et al. (2018), and reported separately on the *wh*-effect in the +Gap (i.e. difference between item (a) and (b)) and −Gap condition (i.e. difference between item (c) and (d)). Within the +Gap condition, in items (a) and (b), networks needed to thread the expectation for a gap over the syntactically complex island configurations, while coming across an illicit gap inside the island. Networks seemed to be successful at that; the *wh*-effect for both island types was decreased in the island condition as compared to the non-island and control conditions. The LSTM networks seemed to struggle within the –Gap conditions, however, in items (c) and (d), in which they needed to thread the expectation for a filled gap over islands, while coming across a licit filled gap inside the island. Here, the *wh*-effect is decreased in both islands compared to non-islands, but not compared to the control condition; none of the LSTM networks are able to recover the filled-gap expectation when moving through a syntactic island configuration with a licit filled gap. LSTM networks can thus successfully suppress and recover the expectation for a gap, but not for a filled gap, which means that they can learn the principle of suppressing and recovering expectations only imperfectly.

3.2.3 Successful but not exceptionless

After the discussed investigations, Wilcox et al. (2021) decided to combine all the knowledge gathered in these studies into the largest investigation to date on the neural network's learning ability of filler-gap dependencies and their island constraints. This investigation uses both the interaction design introduced by Wilcox et al. (2018) and the gendered expectation effect introduced by Wilcox et al. (2019b) to control for any complexity effects.

Wilcox et al. (2021) start by investigating whether all standard characteristics of regular filler-gap dependencies can be learned by ANNs. First, they show that different types of neural networks can learn that filler-gap dependencies are flexible; gaps can be at different sites (e.g. subject and object position). However, they also admit that some could argue that the network might simply learn the linear relationship between fillers and gaps; if it comes across a *wh*-filler, it learns to expect a gap somewhere. Therefore, they also tested an hierarchical constraint dependencies are subjected to, namely the fact that a gap always needs to be in the c-command domain of the filler. For example, in (11a), the gap is in the c-command domain of the filler ‘who’, but in (11b) it is not.

(11) a. [$_{DP}$ The fact that John knows **who$_i$** Mary saw **___$_i$**] surprised them today.
b. *[$_{DP}$ The fact that John knows **who$_i$** Mary saw **Peter**] surprised **___$_i$** today.

Their results showed that neural networks can learn to adhere to these hierarchical constraints; the *wh*-licensing interaction is reduced for sentences as (11b) as compared to sentences such as the one in (11a). However, once again, some could argue that these effects solely arise due to the fact that the c-command domain is linearly closer to the filler. Therefore, Wilcox et al. (2021) also examined whether the neural networks could link a filler to a gap regardless of the number of intervening elements. Different types of neural networks were able to represent filler-gap dependencies, even when the sentence contained five layers of embedding.

To truly show that neural networks can learn filler-gap dependencies, however, the constraints that apply to these dependencies need to be learned as well (Da Costa & Chaves, 2020). Therefore, Wilcox et al. (2021) tested all islands previously tested by Wilcox et al. (2019b), and showed that *wh*-, adjunct, CNP, left branch, and coordinate structure islands could all successfully be learned by different types of neural networks. Moreover, unlike other studies, they also showed successful results for subject islands, but only when the subject was non-sentential. Also important to note is that these results could not be due to processing factors, as the control condition used ruled out this option.

Despite all of these successful results, Chaves (2020) remains sceptical about whether neural networks are capable of learning the syntactic constraints on filler-gap dependencies, for two reasons. First, as found in many of the studies discussed before, (sentential) subject islands seem impossible to learn for the network. According to Chaves, this is striking, because it is one of the strongest island effects. For example, it is regarded as a stronger constraint than the *wh*-island effect, while the latter is successfully learned in all studies. Why are the neural networks not able to learn such a strong, almost exceptionless, constraint? Second, he rightly states that island constraints are very complex. As discussed in Section 2.1, this is also put forward by many linguists in the theoretical debate about islands; many exceptions are found, making it difficult to create a constraint that works for all possible examples. Chaves argues that, before we can state that neural networks can learn island constraints, it first has to be shown that all of the exceptions can be learned as well.

Using the interaction design introduced by Wilcox et al. (2018), Chaves (2020) thus investigated whether neural networks can also learn the exceptions to the island constraints. The first exception concerns CNP islands, which arguably disappear in existential relative clauses, illustrated in (12).

(12) Which diamond ring$_i$ did you say there was [$_{CNP}$ nobody in the world who could buy ___$_i$]?

(Chaves, 2020, p. 24)

While Wilcox et al. (2018) originally hypothesized that the full *wh*-licensing interaction should decrease for islands in comparison to non-islands, it should thus not decrease for the island in (12) as this is argued to be an acceptable sentence. Only one of two LSTM networks did not show this decrease.

Second, Chaves (2020) tested the network on conditional adjunct islands, judged as highly acceptable in Chaves and Putnam (2020), an example of which is shown in (13).

(13) What$_i$ does Evan get grumpy [$_{\text{conditional adjunct}}$ if he is told to do ___$_i$]?

(Chaves, 2020, p. 25)

While the full *wh*-licensing interaction should thus not decrease for this conditional adjunct island as these are not considered gap-resistant by humans, it did for both LSTM networks.

Last, Chaves (2020) examined the learnability of negative phrase islands as these can be affected by several semantic and pragmatic factors. For example, the sentence in (14a) is unacceptable because of semantic reasons; using an existential modal makes the island effect disappear, as is shown in (14b). The two LSTM networks did not show any sensitivity to these factors, however.

(14) a. *How fast$_i$ didn’t John drive ___$_i$?
b. How fast$_i$ is John required not to drive ___$_i$?

(Chaves, 2020, p. 27)

Therefore, Chaves (2020) concludes that island constraints are too complex for neural networks to learn completely. He argues that this is because “filler-gap dependencies (…) involve rich morphological, syntactic and semantic dependencies which crucially interact with pragmatics and world knowledge” (p. 28), making them impossible to learn from training data alone.

In sum, a uniform conclusion about whether neural networks are able to learn syntactic island constraints does not exist (yet) as previous investigations show different results. Warstadt et al. (2019) thus rightly state that syntactic island constraints are the hardest phenomenon to learn for RNNs. While these networks can learn to represent regular filler-gap dependencies, they still struggle to learn the different structure types in which these dependencies cannot be made. Therefore, more research is necessary in English on the island constraints not successfully learned yet. For the successfully learned island constraints, however, it can already be investigated whether they can also be successfully learned in languages other than English. The *wh*- and coordinate structure island constraints are, for example, successfully learned in various studies in English (e.g. Wilcox et al., 2018; Wilcox et al., 2019b; Wilcox et al., 2021), making it interesting to see whether this success is limited to the English language only or whether it can also be achieved in other languages. Therefore, the current research project specifically focused on the successfully learned *wh*- and coordinate structure island constraints.

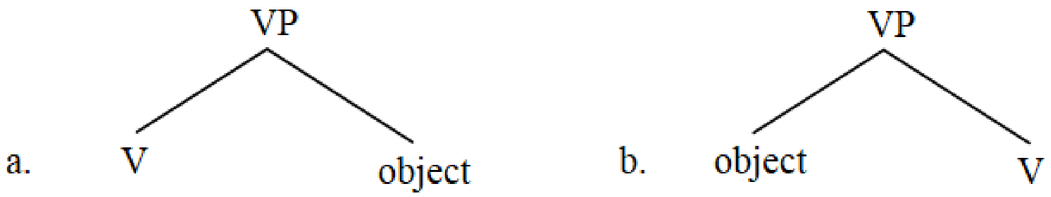
Important to note is that all results and conclusions discussed in the current section are limited to the English language. The learnability of syntactic island constraints has not been investigated in another language yet, while island effects do exist in languages all over the world (Wilcox et al., 2021). To be able to state that the LSTM network can learn syntactic island constraints, and to be able to make any statements about the need for an innate language ability based on these results, more languages need to be investigated. Therefore, the current research project will investigate the learnability of the *wh*- and coordinate structure constraints in Dutch.

3.3 Concerns for research with artificial neural networks

While various investigations have taken place on the learnability of syntactic island constraints by artificial neural networks, most, if not all, have been performed in English. This is concerning as recent literature suggests that these networks bear a bias specifically for right-branching structures, and consequently for right-branching languages, such as English (Dyer et al., 2019). A language's branching direction is determined by the order of the head and the complement of a syntactic phrase. For example, in a verb phrase (i.e. VP), the verb appears in the head V and the object occurs in its complement. This complement can either precede or follow the head, and this order determines a language's branching direction. If the complement always follows the head, as in Figure 5A, the language is referred to as right-branching (e.g. English). Conversely, if the complement always precedes the head, as in Figure 5B, the language is called left-branching (e.g. Japanese). It is also possible to have both right- and left-branching structures, which makes a language mixed (e.g. Dutch) (Frazier & Rayner, 1988).

Figure 5

The structure of the Verb Phrase for right-branching (A) and left-branching (B) languages.



The right-branching structures seem to overlap with a non-linguistic bias in the artificial neural network (Davis & van Schijndel, 2020). While this right-branching bias will therefore inflate the architecture's performance in right-branching languages, it will undermine its performance in left-branching and possibly mixed-branching languages (Li et al., 2020). Therefore, to address this concern, the current study will investigate the learnability of syntactic island constraints not in a right-branching language, but in a mixed-branching language, namely Dutch.

Another possible concern when investigating the syntactic learning abilities of neural networks is that the outputted probability distribution is not comparable with the measurement used to test humans; in this case, acceptability judgements. In the current research project, the performance of the neural networks was assessed by examining the surprisal values that the network assigned to the words in the test sentences. This value represents the extent to which a word is unexpected by the network (Levy, 2008). The network's performance was compared to that of human native speakers, who judged the same test sentences on their acceptability in

Dutch. While previous research has shown that surprisal is indicative of real-time human language processing (Smith & Levy, 2013), and can thus be compared with human reading times, not much research has compared surprisal values with acceptability judgements yet, giving rise to the concern as to whether this is even possible. As stated in Section 2.1, acceptability judgements have been shown to be gradient (Abeillé et al., 2020; Hofmeister & Sag, 2010), which suggests that the knowledge underlying these judgements is probabilistic in nature instead of categorical (Lau et al., 2017; Sprouse et al., 2018). Moreover, multiple previous investigations have argued that acceptability is a concept comparable to probability (Pearl & Sprouse, 2015; Wilcox et al., 2021). Pearl and Sprouse (2015) even suggests a way in which the probabilities assigned to words and sentences can be translated to acceptability judgements; higher surprisal values, and thus less probable words and sentences, link to lower judgements, and lower surprisal values, and thus more probable words and sentences, to higher judgements.

While there have thus been arguments, there has not been a direct comparison between surprisal values and acceptability judgements on a syntactic phenomenon that has been shown to be learnable by ANNs in different languages. This direct comparison could show whether a network's probability distribution and humans' acceptability judgements are indeed comparable. Based on previous research, however, I have compared the two in the current research project, using Pearl and Sprouse's (2015) translation described above.

4. The current research

The current research project investigated whether artificial neural networks can learn to be sensitive to the *wh*- and coordinate structure island constraints in Dutch, comparable to human native speakers of Dutch. To answer this research question, it first needs to be established whether the *wh*- and coordinate structure island constraints exist in Dutch, and if so, to what extent human native speakers are sensitive to these constraints. Therefore, an acceptability judgement task was performed using the interaction design introduced by Wilcox et al. (2018). For each island type, the experimental items were manipulated for PRESENCE OF GAP (no gap vs. gap), PRESENCE OF FILLER (no filler vs. filler) and ISLAND (non-island vs. island). Example sentences of all these conditions can be found in Table 6 and Table 7. The native Dutch speakers had to judge on a 7-point scale whether the regular filler-gap dependencies, *wh*-islands and coordinate structure islands with and without gaps and/or fillers were acceptable in Dutch.

Moreover, following Wilcox et al. (2019b), control items were added with gender expectations to control for the complexity effect discussed in Section 3.3.1. Similar to Wilcox et al. (2019b), these control items were manipulated for GENDER MATCH (match vs. no match) and STRUCTURE (non-island vs. *wh*-island vs. coordinate structure island).

After the human native speakers of Dutch judged these items, a LSTM network, trained on 12 million sentences extracted from the Dutch *Corpora Of the Web* (NLCOW14), was presented with the same test sentences. I examined the surprisal values it assigned to each word of the test sentences, which indicate the extent to which a word was unexpected by the network, to assess its sensitivity to the island constraints. These surprisal values were then compared to the acceptability judgements using the translation proposed by Pearl and Sprouse

(2015); the higher the surprisal value, and thus the more unexpected a sentence (part) was, the lower the acceptability rating. I will now discuss the hypotheses set beforehand.

4.1 Hypotheses

I will now discuss the hypotheses for the acceptability judgement and the modelling task. After discussing the predictions made for these tasks individually, I will also hypothesize about the comparison between the Dutch native speakers and the LSTM network.

For both the human native speakers and the LSTM network, based on Beljon et al. (2021) and unpublished results (Suijkerbuijk, 2021), it is expected that they are sensitive to *wh*-island violations, showing that the *wh*-island constraint exists in Dutch. Moreover, as extraction of (part of) a full conjunct does not seem possible in any language (Liu et al., 2022), the coordinate structure island constraint is also expected to exist in Dutch. Consequently, the following is predicted for the interaction design by Wilcox et al. (2018).

First, as previous research has shown that humans and neural networks may simply not be able to thread information through syntactically complex constructions (i.e. islands) (Keshev & Meltzer-Asscher, 2019; Chowdhury & Zamparelli, 2018), control items were added, examples of which can be found in Table 8. It is predicted that the sentences in which the gender of the noun phrase matches the gender of the possessive pronoun will be judged as more acceptable and will be less unexpected than sentences in which there is no match. Moreover, if the native speakers and the LSTM network have trouble threading information through island configurations, an interaction is expected between GENDER MATCH and STRUCTURE; the gendered expectation effect, i.e. the difference between the sentences with matching and non-matching genders, will reduce within island configurations. However, if the native speakers and the LSTM network can work within complex structures, no interaction effect is expected to arise, meaning that the gendered expectation effect will not be different between non-islands and the two island types.

Second, the filler-gap dependencies and island constraints will be examined. As discussed before, the interaction design by Wilcox et al. (2018) was based on two assumptions: (1) gaps require fillers, and (2) fillers require gaps. If Dutch speakers and the LSTM network indeed assume the first, gaps should be more surprising and less acceptable when no *wh*-filler is present. For the sentences in Table 6 and Table 7, this means that the second sentence should be more surprising and less acceptable than the first. Furthermore, if the native speakers and the network also follow the second assumption, filled argument positions should be more surprising and less acceptable when a *wh*-filler is present. This means that the fourth sentence should be less surprising and more acceptable than the third.

However, when the gaps and fillers appear in island configurations, the hypotheses are different. When the island configuration contains a gap, the presence of a *wh*-filler should not affect the network's expectations or the speaker's acceptability. Regardless of the presence of a filler, the network should never expect a gap inside an island and a Dutch speaker should never find a gap inside an island acceptable. However, unlike the hypothesis put forward by Wilcox et al. (2018), when there is no gap but a filler argument position inside the island, I expect the presence of a filler to affect the network's expectations and the speakers' acceptability judgements. While the network should never expect a gap inside an island, coming across a *wh*-filler at the start of the sentence should give rise to the expectation of a gap somewhere else. When encountering the period whilst not having encountered a gap, there should be a spike in surprisal, as this sentence is completely ungrammatical. The native speakers will thus also find it unacceptable to come across a *wh*-filler but no gap. This all means that a three-way interaction effect should arise between PRESENCE OF GAP, PRESENCE OF FILLER, and ISLAND, as the interaction between the former two differs between non-islands and islands.

Moreover, two additional predictions were made. First, the current research also directly compared the two island types investigated by adding STRUCTURE (*wh*-island vs. coordinate structure island) to the design. While previous research claims that coordinate structure island effects occur in every language (Liu et al., 2021), it varies between and within languages whether sensitivity to *wh*-island effects is shown (Pañeda et al., 2020). Therefore, while I can be sure to find a strong coordinate structure island effect, it is uncertain whether I will find a *wh*-effect similar in strength to that found in Beljon et al. (2021) for Dutch. Therefore, it is predicted that a main effect of STRUCTURE could occur; coordinate structure islands could be judged as less acceptable, and the network could be more surprised to see coordinate structure islands than it will be to come across *wh*-islands. Moreover, the interaction effect discussed before between PRESENCE OF GAP, PRESENCE OF FILLER, and ISLAND will be more pronounced for coordinate structure islands as compared to *wh*-islands, resulting in a four-way interaction between PRESENCE OF GAP, PRESENCE OF FILLER, ISLAND, and STRUCTURE. Second, previous literature has shown a learning effect for the ungrammatical island configurations; the acceptability of these islands increased after participants were more often exposed to them (Christensen et al., 2013; Kush et al., 2019). Therefore, for the analysis of the acceptability judgement task, TRIAL PROGRESS was also added, and predicted to have an effect on the interaction between PRESENCE OF GAP, PRESENCE OF FILLER and ISLAND; the acceptability of the ungrammatical islands [+GAP, +FILLER, +ISLAND], [+GAP, -FILLER, +ISLAND] and [−GAP, +FILLER, +ISLAND] was predicted to increase, the more the participants progressed through the experiment.

After conducting both the acceptability judgement and the modelling task, the performance of the native Dutch speakers and the LSTM network will be compared to see whether the network shows a sensitivity to the *wh*- and coordinate structure island violations similar to that observed by the native speakers. In general, based on the success of the neural network on island constraints in English (e.g. Wilcox et al., 2021), I expect the LSTM network to show roughly the same sensitivity to the *wh*- and the coordinate structure island violations in Dutch as the human native speakers. This means that any effects observed for the native speakers will also be observed for the LSTM network; for example, if the results of the judgement task show a three-way interaction effect between PRESENCE OF GAP, PRESENCE OF FILLER and ISLAND, this interaction effect will also arise in the modelling task. Moreover, when looking into the statistical effects, the acceptability judgements and the surprisal values will show the same pattern; for example, if the presence of a filler decreases the acceptability of sentences with a gap, it will increase the surprisal values of these sentences assigned by the network. By comparing the human native speakers and the LSTM network on the hypotheses described above, it can be determined whether the performance by the LSTM network is actually human-like.

5. Human acceptability judgement task

5.1 Methodology

To test whether the *wh*-island and the coordinate structure island constraints exist in Dutch and, if so, to what extent native speakers are sensitive to them, an acceptability judgement task using a 7-point Likert scale was conducted in *Qualtrics* (Qualtrics, March 2022). This research project was approved by the ethics assessment committee for the humanities of Radboud University (ETC-GW number 2022-0232).

5.1.1 Experimental design

This experiment contained experimental and control items, for which a different experimental design was used. For the experimental items, four independent variables were included:

1. PRESENCE OF GAP: is a gap present in the sentence? This independent variable had two levels (NO GAP VS. GAP) and was measured within-subjects and within-items;
2. PRESENCE OF FILLER: is a *wh*-filler present in the sentence? This independent variable had two levels (NO FILLER VS. FILLER) and was measured within-subjects and within-items;
3. ISLAND: does the sentence contain an island or a regular filler-gap dependency? This independent variable had two levels (NO ISLAND VS. ISLAND) and was measured within-subjects and within-items;
4. STRUCTURE: what type of island does the sentence contain? This independent variable had two levels (WH-ISLAND VS. COORDINATE STRUCTURE ISLAND) and was measured within-subjects and between-items.

The inclusion of these independent variables resulted in 16 different conditions in total. This means that there were eight conditions per island type. Example sentences of these eight conditions for *wh*-islands can be found in Table 6, and in Table 7 for coordinate structure islands.

For the control items, two independent variables were included:

1. GENDER MATCH: does the gender of the NP in the matrix and the embedded sentence match? This independent variable had two levels (NO MATCH VS. MATCH) and was measured within-subjects and within-items;
2. STRUCTURE: is the sentence a regular filler-gap dependency or does it contain an island violation, and if so, which type? This independent variable had three levels (NO ISLAND VS. WH-ISLAND VS. COORDINATE STRUCTURE ISLAND) and was measured within-subjects and within-items.

Consequently, there were six control conditions in total. Example sentences of these conditions can be found in Table 8. The standardized judgements per participant were used as the dependent variable in the analysis of both the experimental and the control items.

Table 6

Example sentences for the experimental items with a wh-*island.*

Gap?	Filler?	Island?	Example sentence
Yes	Yes	No	*Ik weet wat jij denkt dat de bakker maakt in de bakkerij.* I know what you think that the baker makes in the bakery ‘I know what you think that the baker makes in the bakery.’
Yes	No	No	*Ik weet dat jij denkt dat de bakker maakt in de bakkerij.* I know that you think that the baker makes in the bakery ‘I know that you think that the baker makes in the bakery.’
No	Yes	No	*Ik weet wat jij denkt dat de bakker koekjes maakt in de bakkerij.* I know what you think that the baker cookies makes in the bakery ‘I know what you think that the baker makes cookies in the bakery.’
No	No	No	*Ik weet dat jij denkt dat de bakker koekjes maakt in de bakkerij.* I know that you think that the baker cookies makes in the bakery ‘I know that you think that the baker makes cookies in the bakery.’
Yes	Yes	Yes	*Ik weet wat jij je afvraagt of de bakker maakt in de bakkerij.* I know what you REF wonder whether the baker makes in the bakery ‘I know what you wonder whether the baker makes in the bakery.’
Yes	No	Yes	*Ik weet dat jij je afvraagt of de bakker maakt in de bakkerij.* I know that you REF wonder whether the baker makes in the bakery ‘I know that you wonder whether the baker makes in the bakery.’
No	Yes	Yes	*Ik weet wat jij je afvraagt of* I know what you REF wonder whether *de bakker koekjes maakt in de bakkerij.* the baker cookies makes in the bakery ‘I know what you wonder whether the baker makes cookies in the bakery.’
No	No	Yes	*Ik weet dat jij je afvraagt of* I know that you REF wonder whether *de bakker koekjes maakt in de bakkerij.* the baker cookies makes in the bakery ‘I know that you wonder whether the baker makes cookies in the bakery.’

Table 7

Example sentences for the experimental items with a coordinate structure island.

Gap?	Filler?	Island?	Example sentence
Yes	Yes	No	*Ik weet wat jij denkt dat de man aanbiedt tijdens de veiling.* I know what you think that the man offers during the auction ‘I know what you think that the man offers during the auction.’
Yes	No	No	*Ik weet dat jij denkt dat de man aanbiedt tijdens de veiling.* I know that you think that the man offers during the auction ‘I know that you think that the man offers during the auction.’
No	Yes	No	*Ik weet wat jij denkt dat de man* I know what you think that the man *het schilderij aanbiedt tijdens de veiling.* the painting offers during the auction ‘I know what you think that the man offers the painting during the auction.’
No	No	No	*Ik weet dat jij denkt dat de man* I know that you think that the man *het schilderij aanbiedt tijdens de veiling.* the painting offers during the auction ‘I know that you think that the man offers the painting during the auction.’
Yes	Yes	Yes	*Ik weet wat jij denkt dat de man* I know what you think that the man *het schilderij en aanbiedt tijdens de veiling.* the painting and offers during the auction ‘I know what you think that the man offers the painting and during the auction.’
Yes	No	Yes	*Ik weet dat jij denkt dat de man* I know that you think that the man *het schilderij en aanbiedt tijdens de veiling.* the painting and offers during the auction ‘I know that you think that the man offers the painting and during the auction.’
No	Yes	Yes	*Ik weet wat jij denkt dat de man* I know what you think that the man *het schilderij en het juweel aanbiedt tijdens de veiling.* the painting and the jewel offers during the auction ‘I know what you think that the man offers the painting and the jewel during the auction.’
No	No	Yes	*Ik weet dat jij denkt dat de man* I know that you think that the man *het schilderij en het juweel aanbiedt tijdens de veiling.* the painting and the jewel offers during the auction ‘I know that you think that the man offers the painting and the jewel during the auction.’

Table 8

Example sentences for the control items.

Match?	Structure?	Example sentence
Yes	No island	*Ik weet dat de meester denkt dat* I know that the teacher.MASC thinks that *de leerlingen zijn uitleg begrijpen.* the students his explanation understand 'I know that the teacher thinks that the students understand his explanation.'
No	No island	*Ik weet dat de meester denkt dat* I know that the teacher.MASC thinks that *de leerlingen haar uitleg begrijpen.* the students her explanation understand 'I know that the teacher thinks that the students understand her explanation.'
Yes	*Wh*-island	*Ik weet dat de meester zich afvraagt of* I know that the teacher.MASC REF wonders whether *de leerlingen zijn uitleg begrijpen.* the students his explanation understand 'I know that the teacher wonders whether the students understand his explanation.'
No	*Wh*-island	*Ik weet dat de meester zich afvraagt of* I know that the teacher.MASC REF wonders whether *de leerlingen haar uitleg begrijpen.* the students her explanation understand 'I know that the teacher wonders whether the students understand her explanation.'
Yes	Coordinate structure Island	*Ik weet dat de meester denkt dat* I know that the teacher.MASC thinks that *de leerlingen zijn uitleg en de sommen begrijpen.* the students his explanation and the sums understand 'I know that the teacher thinks that the students understand his explanation and the calculations.'
No	Coordinate structure island	*Ik weet dat de meester denkt dat* I know that the teacher.MASC thinks that *de leerlingen haar uitleg en de sommen begrijpen.* the students her explanation and the sums understand 'I know that the teacher thinks that the students understand her explanation and the calculations.'

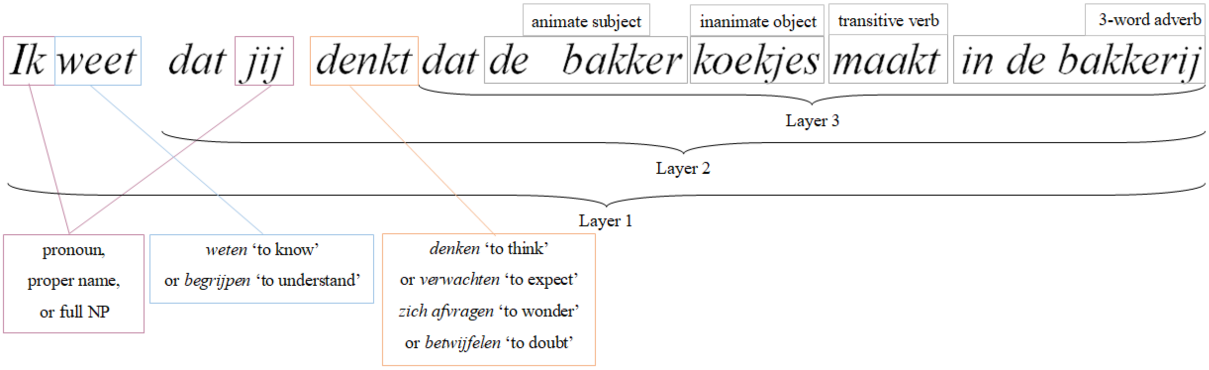
5.1.2 Materials

In the creation of all sentences used in the experiment, it was closely monitored whether the words used for the sentences also occurred in the 20,000 most frequent word list of the corpus used to train the neural network on (described in Section 6.2.1).

First, 32 item sets were made for each island type, each containing the eight experimental conditions (see Table 6 for an example item set for *wh*-islands and Table 7 for an example item set for coordinate structure islands). The sentences in these item sets all roughly have the same syntactic structure, as illustrated for the [NO GAP, NO FILLER, NO ISLAND, WH-ISLAND] condition in Figure 6.

Figure 6

The general syntactic structure of an experimental item, specifically the [no gap, no filler, no island, wh-*island] condition.*

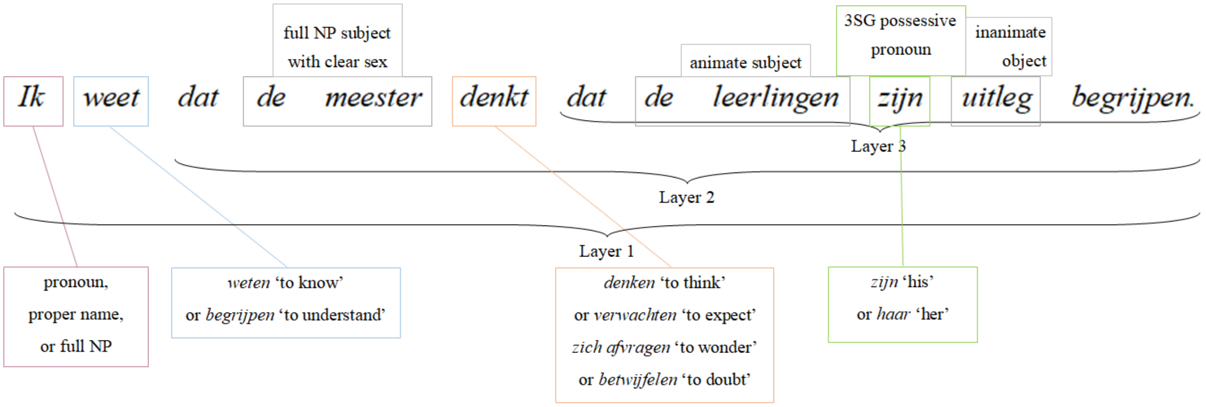

animate subject
inanimate object
transitive verb
3-word adverb
Ik weet dat jij denkt dat de bakker koekjes maakt in de bakkerij
Layer 3
Layer 2
Layer 1
pronoun, proper name, or full NP
weten 'to know' or *begrijpen* 'to understand'
denken 'to think' or *verwachten* 'to expect' *zich afvragen* 'to wonder' or *betwijfelen* 'to doubt'

First, every sentence consisted of three layers of embedding and ended with a three-word long adverbial phrase, e.g. *in de bakkerij* 'at the bakery'. Second, within the third layer, the object (e.g. *koekjes* 'cookies') or the object filler (i.e. *wat* 'what') was always inanimate and the subject always animate (e.g. *de bakker* 'the baker') to minimize subject-object ambiguity (Beljon et al., 2021). Moreover, a verb was used that could only be interpreted in a transitive manner (e.g. *maken* 'to make'). Last, within the first and the second layer, the subjects were either a pronoun, a proper name or a full NP, and two different verbs were alternated (e.g. *weten* 'to know' and *begrijpen* 'to understand' in the first layer; *denken* 'to think' and *verwachten* 'to expect' for non-islands in the second layer; *zich afvragen* 'to wonder' and *betwijfelen* 'to doubt' for islands in the second layer). All the subjects, objects and verbs used in the sentences, and the completed item sets can be found on the OSF-page of this research project (https://osf.io/kt3he/).

In addition to these experimental item sets, 32 control item sets were constructed (see Table 8 for an example item set). Similarly, the control items all roughly had the same syntactic structure, as illustrated in Figure 7 for the [MATCH, NO-ISLAND] condition.

Figure 7

Illustrating the general syntactic structure of a control item, specifically the [match, non-island] condition.


full NP subject with clear sex
3SG possessive pronoun
animate subject
inanimate object
Ik weet dat de meester denkt dat de leerlingen zijn uitleg begrijpen.
Layer 3
Layer 2
Layer 1
pronoun, proper name, or full NP
weten 'to know' or begrijpen 'to understand'
denken 'to think' or verwachten 'to expect' zich afvragen 'to wonder' or betwijfelen 'to doubt'
zijn 'his' or haar 'her'

Similar to the experimental items, each sentence had three layers of embedding, an animate subject and inanimate object in the third layer, the subject in the first layer being either a pronoun, a proper name or a full NP, and two different options for the verb in the first and second layer. In addition, within the third layer, the object was always modified by a third person singular possessive pronoun, either masculine (i.e. *zijn* 'his') or feminine (*haar* 'her'). Last, within the second layer, the subject was always a full NP with an unambiguous semantic gender, either unambiguously masculine (e.g. *meester* 'male teacher') or unambiguously feminine (*juffrouw* 'female teacher'). The gender of the full NP in the second layer and that of the possessive pronoun in the third layer could either match or not, depending on whether the sentence was a mismatch or a match condition of GENDER MATCH. Gender was counterbalanced within match and mismatch; 16 matches were female and 16 matches were male. This also means that 16 of the mismatches were female and the other 16 male.

Besides the experimental and control item sets, 64 filler items were constructed as well based on those by Beljon et al. (2021) and Kovač and Schoenmakers (2022), covering the full range of acceptability: 21 acceptable, 22 moderately acceptable, and 21 unacceptable sentences. The 21 acceptable fillers consisted of regular declarative statements and declarative statements with *gaan* 'to go' as its main verb. The 22 moderately acceptable sentences can also be referred to as marked grammatical sentences, meaning that these contained a grammatical error not generally considered a serious one (Kovač & Schoenmakers, 2022). These sentences included variation in the order of the verb cluster, anglicisms, or violations of the ANIMATE FIRST principle. Last, the 21 unacceptable fillers consisted of subject-verb agreement errors and word salads. These filler items were identical across the experimental lists, and each list started with at least three filler items (one acceptable and two moderately acceptable items). Examples of each of the different filler categories can be found in (15) below.

(15) Regular declarative statement

a. *Jij ziet dat zij vermoedt dat Anne en Tom de boekenkast*
you see that she suspects that Anne and Tom the book.case
aangeschaft hebben voor hun woning.
purchased have for their place
'You see that she suspects that Anne and Tom purchased the book case for their place.'

Declarative statement with *gaan*

b. *Ik zie dat hij verwacht dat de oppas Samir een verhaaltje*
I see that he expects that the babysitter Samir a story
voor gaat lezen.
for goes read
'I see that he expects the nanny to read Samir a story.'

Variation in verb cluster order

c. *Hij beseft dat ik betwijfel of Bas het bord afgewassen hebben zal.*
he realizes that I doubt whether Bas the plate washed have will
'He realizes that I doubt Bas will have washed the plate.'

Anglicism

d. *Hij gelooft dat ik verwacht dat Daphne overnacht in het ziekenhuis*
he believes that I expect that Daphne overnight in the hospital
moet blijven.
must stay
'He believes that I expect Daphne to stay in the hospital overnight.'

Violation of ANIMATE FIRST principle

e. *Zij gelooft dat jij vermoedt dat het theaterstuk de bejaarden*
she believes that you suspects that the play the elderly
tot huilens toe ontroerde.
to crying to moved
'She believes that you suspect that the play moved the elderly to tears.'

Agreement error

f. **Hij ziet dat jij verwacht dat Willem boodschappen gedaan hebben.*
he sees that you expect that Willem groceries done have
*'He sees that you expect that Willem has done the grocery shopping.'

Word salad

g. **Beseffen verwacht jij dat wij Jolie het geld uitgegeven hebben.*
realize expect you that we Jolie the money spent have
*'Realize expect you that we Jolie the money spent have.'

Similar to the experimental and control items, the filler items all also had the same syntactic structure; each item had three layers of embedding, and within the first and second layer, the subject was either a pronoun, a proper name or a full NP and only four different verbs were alternated. However, these verbs differed from those used in the experimental and control items. The filler items were added for two reasons: (1) to ensure that the full range of acceptability is covered by the items, and (2) to check, specifically with the unacceptable items, whether the participant performed the task correctly and attentively (Beljon et al., 2021; Sprouse, 2018).

Only one item of each experimental item set was shown to the participant. As there are eight experimental conditions per island type, this means that eight main lists were made using a Latin square design. Subsequently, the control items were added with a Latin square design and all of the filler items, resulting in a total of 160 items per list. Based on each of these eight main lists, 15 different lists were created, all with a different item order, resulting in 120 lists in total. In the randomization of the item order, one criterion was used: ungrammatical experimental items (i.e. the [NO GAP, FILLER] and [GAP, NO FILLER] conditions for non-islands, and the [NO GAP, FILLER], [GAP, NO FILLER] and [GAP, FILLER] conditions for islands) and ungrammatical control items (i.e. gender mismatches) had to be separated by at least one item. Participants were randomly assigned to one these 120 lists, such that each list was assigned at least once.

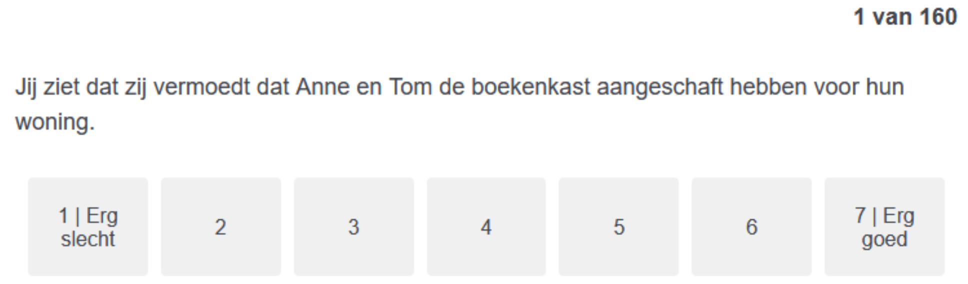
5.1.3 Procedure

The web-based experiment was built using the Qualtrics survey software (Qualtrics, March 2022). First, participants read general information about the university's policy regarding processing of the data, their rights, and my contact details. After they provided consent, they answered demographic questions, concerning their age, gender and the presence of any language/reading disorders. Participants were not able to start the experiment if they were younger than 16 years old or had any language/reading disorders. Participants that did start the experiment were randomly assigned to one of the 120 lists within the experiment.

The participants were presented with 160 sentences (one at a time) and were instructed to imagine that these were produced by a native speaker of Dutch that they know well, e.g. a close friend. They were then told to judge these sentences on how good they sound in Dutch (specifically *hoe goed vindt u de zin klinken?* 'how good do you think the sentence sounds?') on a scale ranging from 1 (*Erg slecht* 'very bad') to 7 (*Erg goed* 'very good'), and to base their judgement on their first intuition. Each participant started with three filler items to familiarise them with the task. An example of an item as presented to the participants can be found in Figure 8. As soon as the participants finished the experiment, they were asked to guess its purpose. No participant guessed that correctly, however. The experiment lasted 20 to 30 minutes.

Figure 8

An example of an item as presented to the participants. The sentence can be translated to 'you see that she suspects that Anne and Tom purchased the bookcase for their home'.



1 van 160

Jij ziet dat zij vermoedt dat Anne en Tom de boekenkast aangeschaft hebben voor hun woning.

5.1.4 Participants

First, 151 native speakers of Dutch entered the experiment. These participants were recruited through the Radboud Research Participation System Sona, and all volunteered to participate and provided consent before entering the experiment. However, 16 of these participants did not complete the experiment, five indicated to have a reading or language disorder, and 42 rated more than two highly ungrammatical filler items a 3 or higher indicating lack of attention to the experiment. Consequently, these 63 participants were excluded. The remaining 88 participants, of which 75 were female, had a mean age of 19 years (SD_{age} = 2.04; range: 17-33).

5.1.5 Data analysis

Performance on the unacceptable filler items was checked before data analysis to determine whether participants had completed the experiment correctly and attentively. Before the start of the experiment, specific exclusion criteria were set, one being that the data of the participant would be excluded if that participant rated more than two of the ungrammatical filler items with word salads or more than two of the ungrammatical filler items with agreement errors with a rating of 3 or higher. After checking the data, this meant that 12 participants had to be excluded based on their performance on the word salads, and that an additional 56 had to be excluded based on their performance on both the word salads and the agreement errors. As this came down to a remarkably high number of excluded participants (68 out of 151 in total), I performed an item analysis for both filler categories (see Appendix A). This showed that all word salads were rated quite low on average (1.30), but that ratings of the agreement errors were higher than expected beforehand (between 2.00 and 3.00). Therefore, the boundary of a rating of 3 of higher set for agreement errors before the start of the experiment turned out to be too low; a rating of 3 actually appeared to be plausible for agreement errors. Consequently, I decided to raise the boundary to 4, which changed the exclusion criterion to the following: “the data of the participant will be excluded if that participant rated more than two word salads or more than two agreement errors with a rating of 4 of higher”. Based on this criterion, 42 participants had to be excluded from further analysis.

Before the statistical analysis, the raw acceptability judgement scores (of experimental, control and filler items) were converted to *z*-scores per participant using all items, to correct for individual differences in scale use. For the statistical analysis, I fitted two linear mixed-effects (LME) models, one for the experimental items and one for the control items. First, for the experimental items, I fitted an LME model to the standardized scores with PRESENCE OF GAP, PRESENCE OF FILLER, STRUCTURE, ISLAND and their interaction as fixed factors, using the *lmer* function from the *lmerTest* package (Kuznetsova et al., 2017) in R (version 3.6.0; R Core Team, 2019). Moreover, the interaction between TRIAL PROGRESS (measured between 0 and 1), PRESENCE OF GAP, PRESENCE OF FILLER and ISLAND was added to control for the effect of when an item was seen, as it is predicted that participants’ acceptability of island violations increase due to more exposure (Christensen et al., 2013; Kush et al., 2019). Second, for the control items, I fitted an LME model to the standardized scores with GENDER MATCH, STRUCTURE and their interaction as fixed effects. The random effects for both models were based on the minimal Akaike Information Criterion (AIC). In addition, all independent variables were coded using simple contrasts. With simple contrasts, the reference level is always coded as -1/3 or -1/2, and the level that is contrasted as 2/3 or ½. For three-leveled variables, the non-contrasted levels are coded as -1/3 as well. The coding scheme for the simple contrasts used is illustrated in Table 9.

Table 9

Coding scheme for the simple contrasts of all independent variables.

		Coding	
	Independent variable	-1/2	1/2
Analysis of experimental items	Presence of gap	No gap	Gap
	Presence of filler	No filler	Filler
	Structure	Coordinate structure island	*Wh*-island
	Island	Non-island	Island
Analysis of control items	Gender match	Match	No match
		Coding	
		-1/3	Contrasted: 2/3 Non-contrasted: -1/3
	Structure	Non-island	Coordinate structure island *Wh*-island

Before fitting the models, a box-cox transformation was performed on the standardized scores (with λ = -.55 for the experimental model, and λ = 1.80 for the control model) so that the transformed data was as close to normally distributed as possible. After fitting this model, model assumptions were checked, which showed that the fitted values and the size of the error (ε) correlated for both models (see Appendix B). Significance values for the coefficients from the two models were calculated using the Satterthwaite approximation in *lmerTest* (Kuznetsova et al., 2017). The interaction effects were further examined using contrasts from the *emmeans* package (Lenth, 2022) in R.

5.2 Results

5.2.1 Analysis of the experimental items

Table 10 shows the mean unstandardized acceptability judgements per condition, and Table 11 the mean standardized acceptability judgements, which are also illustrated for both islands in Figure 9.

Table 10

Mean unstandardized acceptability judgement scores. Associated standard deviations are given between parentheses.

		Non-island		Island	
Gap?	Filler?	*Wh*-island	Coordinate structure island	*Wh*-island	Coordinate structure island
Yes	Yes	2.64 (1.50)	2.67 (1.55)	1.92 (1.04)	1.93 (.99)
Yes	No	2.28 (1.41)	2.44 (1.34)	2.36 (1.37)	2.33 (1.53)
No	Yes	2.60 (1.50)	2.72 (1.64)	2.59 (1.52)	2.65 (1.50)
No	No	4.74 (1.70)	4.79 (1.66)	4.88 (1.62)	4.71 (1.62)

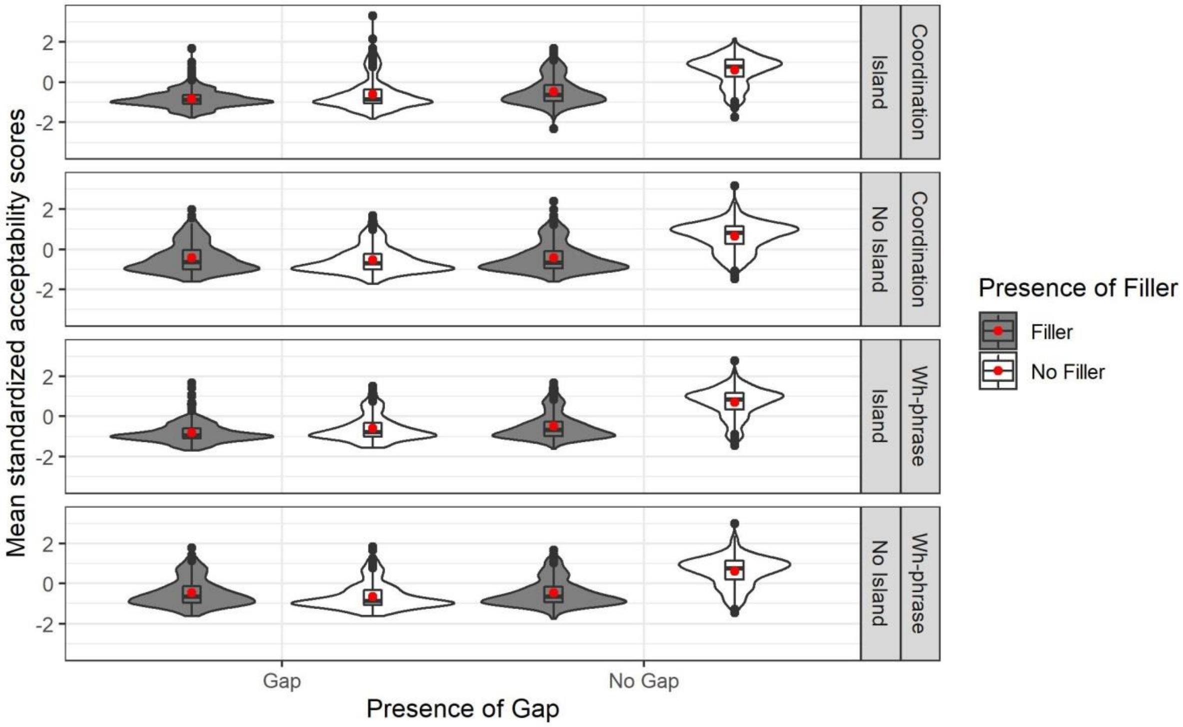
Table 11

*Mean standardized acceptability judgement scores (*z-scores*). Associated standard deviations are given between parentheses.*

		Non-island		Island	
Gap?	Filler?	*Wh*-island	Coordination structure island	*Wh*-island	Coordination structure island
Yes	Yes	-.47 (.68)	-.44 (.73)	-.82 (.50)	-.82 (.46)
Yes	No	-.64 (.65)	-.54 (.66)	-.60 (.64)	-.60 (.78)
No	Yes	-.48 (.67)	-.43 (.72)	-.49 (.67)	-.46 (.67)
No	No	.62 (.74)	.65 (.73)	.70 (.73)	.61 (.73)

Figure 9

Violin/boxplot with the standardized acceptability judgement scores (z-scores) on the y-axis, the levels of PRESENCE OF GAP *(gap vs. no gap) on the x-axis, the levels of* PRESENCE OF FILLER *(filler vs. no filler) representing the different colors, and the levels of* STRUCTURE *and* ISLAND *representing the different boxes. The red dot represents the mean.*



The final model included a random intercept and random slope for FILLER for items, but no random effects for participants. The results of the LME regression analysis are summarized in Table 12.

Table 12

Results of LME analysis of the experimental items

Fixed effects					
Predictors	β [95% CI]	SE	df	*t*	*p*
Trial progress	.02 [.01, .03]	.00	5604	5.16	< .001
Structure	−.00 [−.01, .00]	.00	5604	−.98	.330
Island	−.02 [−.03, −.01]	.00	5604	−4.45	< .001
Presence of filler	−.07 [−.08, −.06]	.01	5604	−12.32	< .001
Presence of gap	−.09 [−.10, −.09]	.00	5604	−20.04	< .001
Trial progress x presence of filler x presence of gap x island	−.01 [−.07, .05]	.03	5604	−.25	.801
Island x structure	.01 [.00, .02]	.00	5604	2.20	.028
Island x presence of filler	−.03 [−.05, −.01]	.01	5604	−3.02	.003
Island x presence of gap	−.01 [−.03, .00]	.01	5604	−1.51	.132
Presence of filler x presence of gap	.14 [.12, .16]	.01	5604	14.61	< .001
Island x presence of filler x presence of gap	−.04 [−.08, −.00]	.02	5604	−2.18	.030

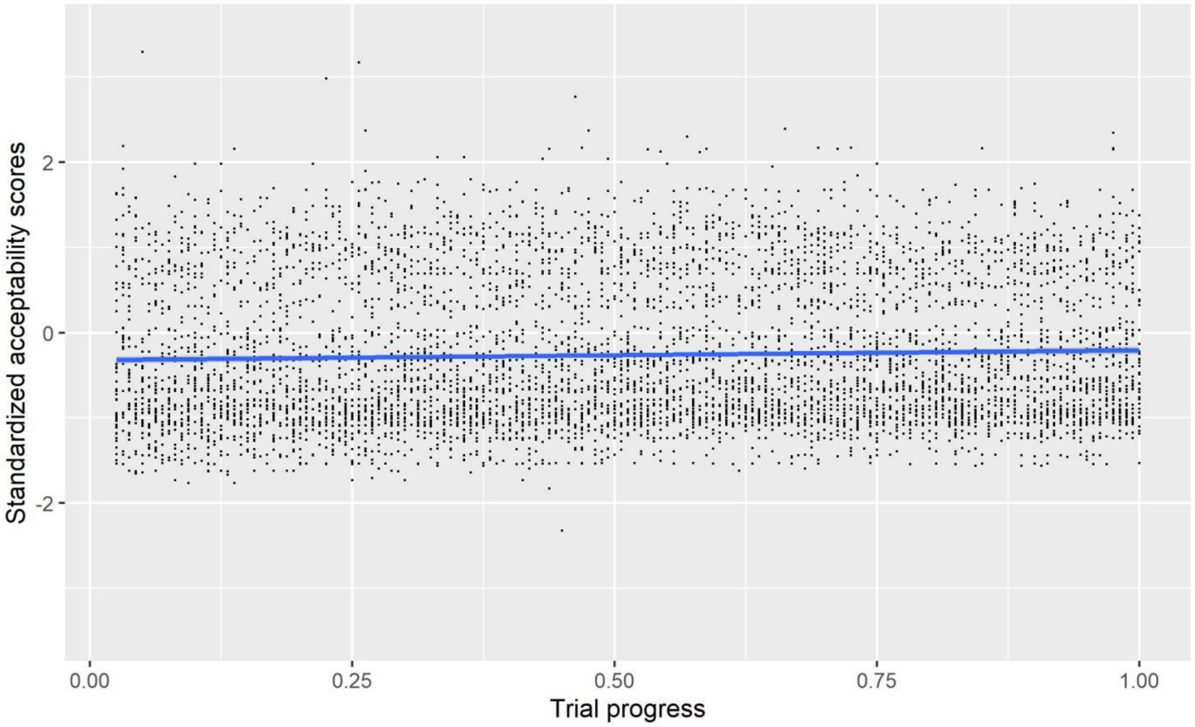
Random effects			
Group	Name	Variance	SD
Item	Intercept	.00	.01
	Filler	.00	.02
Residual		.01	.08

Note. Marginal R^2 = .37; Conditional R^2 = .39; the non-significant interaction effects with STRUCTURE were left out of this table.

First, the analysis revealed a significant main effect of TRIAL NUMBER (β = .02, SE_β = .00, 95% CI of β [.01, .03], p < .001). In general, the acceptability slightly increased as participants were exposed to more items, which is illustrated in Figure 10. However, no interaction effect was found between TRIAL PROGRESS, PRESENCE OF GAP, PRESENCE OF FILLER and ISLAND (β = −.01, SE_β = .03, 95% CI of β [−.07, .05]), indicating that this increase in judgements was not specifically tied to the ungrammatical sentences.

Figure 10

Scatterplot of the main effect of TRIAL NUMBER *with the standardized acceptability judgement scores on the y-axis and the trial progress on the x-axis.*



Furthermore, significant main effects of ISLAND (β = −.02, SE_β = .00, 95% CI of β [−.03, −.01], p < .001), PRESENCE OF FILLER (β = −.07, SE_β = .01, 95% CI of β [−.08, −.06], p < .001) and PRESENCE OF GAP (β = −.09, SE_β = .00, 95% CI of β [−.10, −.09], p < .001) were found, but no main effect of STRUCTURE (β = −.00, SE_β = .00, 95% CI of β [−.01, .00]). This shows that sentences without islands (M = −.22, SD = .85), without fillers (M = .02, SD = .94), and without gaps (M = .09, SD = .90) were rated significantly higher on the 7-point scale than sentences with islands (M = −.31, SD = .87), fillers (M = −.55, SD = .66) and gaps (M = −.62, SD = .66) respectively, but that there was no significant difference between the acceptability ratings of *wh*-islands (M = −.27, SD = .86) and coordinate structure islands (M = −.26, SD = .86).

In addition, a three-way interaction effect between ISLAND, PRESENCE OF FILLER and PRESENCE OF GAP (β = −.04, SE_β = .02, 95% CI of β [−.08, −.00], p = .030) was found. For both non-islands and islands, acceptability decreases when a filler is present and a gap is not ($M_{non\text{-}island}$ = −.45, $SD_{non\text{-}island}$ = .69; M_{island} = −.47, SD_{island} = .67) as opposed to sentences where neither a filler nor a gap is present ($M_{non\text{-}island}$ = .63, $SD_{non\text{-}island}$ = .73; M_{island} = .65,

SD_{island} = .73) (non-island: β = .13, SE_β = .01, p < .001; island: β = .13, SE_β = .01, p < .001). However, non-islands and islands diverge in acceptability on sentences with a gap. While the presence of a filler significantly increases acceptability for non-islands ($M_{+filler}$ = −.46, $SD_{+filler}$ = .70; $M_{-filler}$ = −.59, $SD_{-filler}$ = .66) (β = −.02, SE_β = .01, p = .007), it decreases acceptability in island configurations ($M_{+filler}$ = −.82, $SD_{+filler}$ = .48; $M_{-filler}$ = −.60, $SD_{-filler}$ = .71) (β = .03, SE_β = .01, p < .001).

5.2.2 Analysis of the control items

Table 13 shows the mean unstandardized acceptability judgements per condition. Important to note is that, in general, the control items received higher mean acceptability judgements than the experimental items. The highest rated grammatical condition of the experimental items received a maximum rating of 4.88, while the match-condition of the control items received mean ratings above 5.00. Table 14 shows the standardized scores, which are also displayed in Figure 11.

Table 13

Mean unstandardized acceptability judgement scores by GENDER MATCH *and* STRUCTURE. *Associated standard deviations are given between parentheses.*

Match?	No island	*Wh*-island	Coordinate structure island
Yes	5.45 (1.43)	5.45 (1.45)	5.35 (1.45)
No	3.98 (1.90)	4.06 (1.93)	3.95 (1.88)

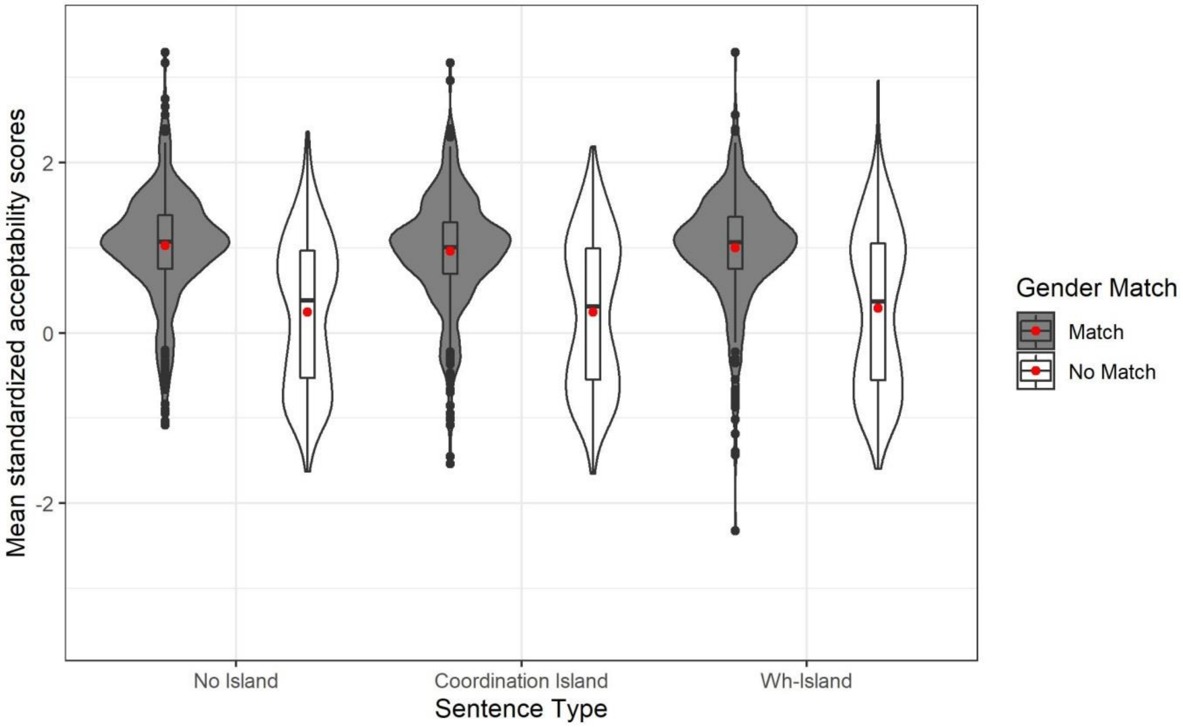
Table 14

Mean standardized acceptability judgement scores (z-scores) by GENDER MATCH *and* STRUCTURE. *Associated standard deviations are given between parentheses.*

Match?	No island	*Wh*-island	Coordinate structure island
Yes	1.02 (.65)	1.00 (.64)	.96 (.66)
No	.24 (.87)	.29 (.94)	.24 (.90)

Figure 11

Violin/boxplot with the standardized acceptability judgement scores (z-scores) on the y-axis, the levels of STRUCTURE *(no island vs. wh-island vs. coordinate structure island) on the x-axis, and the levels of* GENDER MATCH *(match vs. no match) representing the different colors. The red dot represents the mean.*



The final model included a random intercept and random slope for MATCH for items, but no random effects for participants. The results of the LME regression analysis are summarized in Table 15.

Table 15

Results of LME analysis of the control items

Fixed effects

Predictors	β [95% CI]	SE	df	*t*	*p*
Island 1 (no island vs. coordinate structure island)	−.08 [−.31, .15]	.12	2806	−.65	.514
Island 2 (no island vs. *wh*-island)	.03 [−.19, .26]	.12	2806	.30	.767
Match	−2.37 [−2.77, −1.97]	.20	2806	−11.64	< .001
Match x island 1	.27 [−.18, .73]	.23	2806	1.17	.240
Match x island 2	.23 [−.22, .69]	.23	2806	1.00	.317

Random effects

Group	Name	Variance	SD
Item	Intercept	.29	.54
	Match	1.05	1.03
Residual		6.14	2.48

Note. Marginal R^2 = .17; Conditional R^2 = .24.

The analysis revealed a significant main effect of GENDER MATCH (β = −2.37, SE_β = .20, 95% CI of β [−2.77, −1.97], p < .001), indicating that sentences containing a gender mismatch (M = .26, SD = .90) were rated as less acceptable than sentences with a gender match (M = .99, SD = .65). There was no significant main effect of ISLAND (F(2, 2797.08) = .47, p = .628) and no significant interaction effect between GENDER MATCH and ISLAND (F(2, 2796.40) = .81, p = .446), however. Overall, the three island types did thus not differ in acceptability ($M_{non\text{-}island}$ = .63, $SD_{non\text{-}island}$ = .86; $M_{wh\text{-}island}$ = .65, $SD_{wh\text{-}island}$ = .88; $M_{coordination\ island}$ = .60, $SD_{coordination\ island}$ = .86), and the gendered expectation effect, represented by a significant difference between sentences with and without a gender match, did not differ in its magnitude between the three island types either (see Table 13 and Table 14 for means and standard deviations).

5.3 Discussion

To investigate whether the *wh*- and coordinate structure island constraints exist in Dutch, and if so, to what extent human native speakers are sensitive to these constraints, an acceptability judgement task was conducted. The experimental design of this task was based on the interaction design introduced by Wilcox et al. (2018). Moreover, following Wilcox et al. (2019b), control items were added with gender expectations to control for the complexity effect discussed in Section 3.3.1.

Previous research has argued that humans may simply not be able to thread information through syntactically complex constructions such as islands (Keshev & Meltzer-Asscher, 2018). The complexity of island configurations causes processing difficulties, causing the participants to have difficulties maintaining any information in or retrieving it from their working memory. This means that the results of the interaction design could all simply be due to this complexity effect, instead of the speakers actually obeying the syntactic constraints. Consequently, control items were added, which contained a full NP with an unambiguous semantic gender (either masculine or feminine) and a possessive pronoun that could either match the full NP in gender or not.

Overall, it was predicted that sentences with a gender match would be more acceptable than sentences with a mismatch, which will represent a gendered expectation effect. Moreover, if native speakers are not able to thread information through islands, this gendered expectation effect would reduce within island configurations. While the current results showed a gendered expectation effect in general, this effect did not differ in magnitude between non-island and island configurations. This suggests that native speakers seem able to thread a gender expectation through an island configuration, and that they should thus be able to maintain a gap expectation as well when moving through an island. In conclusion, the results of this control study seem to indicate that any result found in the current experiment should not be immediately attributed to processing difficulty caused by complexity.

After conducting the control study, the main experiment was performed, in which the acceptability of *wh*-island and the coordinate structure island violations was tested with Wilcox et al.'s (2018) interaction design. I will first discuss the results on the regular filler-gap dependencies, i.e. the non-islands, and then the results on the island configurations.

Wilcox et al.'s (2018) design was based on two assumptions: (1) gaps require fillers, and (2) fillers require gaps. If native speakers indeed assume the latter, filled argument positions should be less acceptable in non-islands when a *wh*-filler is present. This hypothesis was confirmed by the current results; the presence of a *wh*-filler decreased the acceptability of sentences without gaps and thus with filled argument positions.

Moreover, if humans assume that gaps require fillers, gaps should be less acceptable when no *wh*-filler is present. The results indeed show a significant decrease in acceptability when no filler is present in sentences with a gap, and these sentences are all perceived as unacceptable. Regardless of whether the sentence contains a *wh*-filler, sentences with gaps are judged in the unacceptable range of the 7-point scale. This is remarkable, because sentences with both a gap and a *wh*-filler are perfectly grammatical in Dutch. This result could be due to the complex syntactic structure of the sentences used in the current research project; only bare *wh*-fillers were used and each sentence contained three layers of embedding with the filler in the second and the gap in the third layer. This caused the filler and the gap to be separated by five intervening words in each sentence. This bare filler and the syntactic distance between the filler and the gap can cause difficulties with processing, as they make it harder for the participant to represent, maintain and retrieve the *wh*-filler from their working memory when encountering a gap (Abeillé et al., 2020; Hofmeister & Sag, 2010; Liu et al., 2022; Newmeyer, 2016). Consequently, if a sentence is difficult to process, it is assumed it will receive a lower rating. However, previous research clearly states that filler-gap dependencies are not constrained by the length of the dependencies (Sprouse & Hornstein, 2013), and the control study discussed above showed that participants were able to thread gender expectations through these embedding layers. While this should rule out the processing explanation, Schippers (2012) showed that the acceptability of Dutch long distance *wh*-dependencies are affected by the number of embedding layers; these dependencies were rated as less acceptable when the sentence contained two layers instead of one. Although the number of embedding layers could thus have influenced the acceptability of the grammatical filler-gap dependency, it still shouldn't have decreased it to the same rating that the completely ungrammatical sentences received. Therefore, the ungrammatical rating of the grammatical filler-gap dependency remains difficult to account for. A replication of the current study and more research into the processing of these complex sentences is necessary to unravel the mystery.

Next, the constraints on these regular filler-gap dependencies were tested, namely the island constraints. Unlike the hypothesis put forward by Wilcox et al. (2018), when there is no gap but a filled argument position inside the island, I expected the presence of a filler to affect the acceptability judgement. That is, because it is ungrammatical in Dutch to have a *wh*-filler but no gap. The current results indeed showed that filled argument positions are less acceptable when a *wh*-filler is present.

On the other hand, when the island configuration contains a gap, the presence of a *wh*-filler should not affect its acceptability. Regardless of the presence of a filler, a Dutch speaker should never find a gap inside an island acceptable. While all island configurations with gaps were rated in the ungrammatical range of the 7-point scale, the presence of a filler decreased these ratings even more. This pattern in acceptability can easily be explained, however. Native speakers should not expect a gap within an island, but coming across a *wh*-filler at the start of the sentence should give rise to the expectation of a gap somewhere else, leading to storage of the *wh*-filler in the working memory. When this expectation is not met by encountering a gap somewhere outside of the island, the filler cannot be linked back to a gap, causing the acceptability rating of that sentence to decrease. However, this decrease in acceptability is smaller when there is no *wh*-filler present in the sentence. Without a *wh*-filler, nothing will be stored and maintained in the working memory. Therefore, the only processing cost this sentence has is the presence of a gap inside an island.

Last, additional hypotheses were posed about (1) the difference between *wh*- and coordinate structure islands, and (2) the effect of trial progress on the acceptability of ungrammatical sentences. First, it was predicted that coordinate structure islands would receive lower ratings, and that the interaction effect described before would be more pronounced for coordinate structure islands. This prediction was not borne out, however, as there was no difference in acceptability between the two island types; the current experiment showed that both the *wh*- and the coordinate structure island constraint exist in Dutch, although the *wh*-island effect was less strong than previously found in the literature for Dutch (i.e. Beljon et al., 2021; Suijkerbuijk, 2021). Second, it was predicted that participants' acceptability of the ungrammatical island configurations would increase due to more exposure (Christensen et al., 2013; Kush et al., 2019). In general, the results showed a minimal increase in acceptability as people progressed through the experiment, but this effect was not specifically tied to the ungrammatical islands. The current experiment thus did not show a learning effect specific to syntactic islands.

Now that it is clear that the *wh*- and coordinate structure island constraints exist in Dutch as human native speakers are clearly sensitive to these, it will be investigated to what extent an LSTM network also shows sensitivity to these, and whether this sensitivity is human-like.

6. Testing the artificial neural network

To test whether an artificial neural network, specifically a Long Short-Term Memory (LSTM) model, can learn to be sensitive to the *wh*- and the coordinate structure island constraints in Dutch, it was first trained on sentences extracted from the Dutch *Corpora Of the Web* (NLCOW2014) and then tested on the same sentences that the native speakers judged in the experimental task.

6.1 The Long Short-Term Memory model

The Long Short-Term Memory (LSTM) model (Hochreiter & Schmidhuber, 1997) is a powerful type of recurrent neural network (RNN) that, unlike the traditional Simple Recurrent Network (Elman, 1990), can maintain dependencies over long distances (van Houdt et al., 2020; Wilcox et al., 2021). I chose to employ this network for two reasons, namely (1) it has shown to be successful in previous research on the learnability of syntactic island constraints (e.g. Wilcox et al., 2018, 2019b, 2021), and (2) it possesses traits relevant within the debate about linguistic nativism, specifically it is both domain-general and weakly biased. That the model is a domain-general learner means that it can learn and process any type of input (Frost et al., 2015), and that it can thus generalize about more than just language (Wilcox et al., 2021). Weakly biased entails that the model was not designed specifically for the task of learning language, but that it remains task-general (Lappin & Shieber, 2007). These two traits make the neural network relevant in the debate about innate language abilities/knowledge (Pearl & Sprouse, 2013), in which the current research resides. That is because, with such a model, it is possible to investigate whether a domain-general and weakly biased learner can acquire syntactic knowledge comparable to native speakers, who have been argued to be domain-specific and strongly biased learners (Chomsky, 1971). More information about the specific architecture of the LSTM model will follow in Section 6.2.2.

6.2.Training process

6.2.1 Training data set

The training sentences were extracted from the NLCOW2014 corpus, literally meaning *Corpora from the Web* in Dutch (*NL*), which comprises individual sentences of Dutch texts collected from the World Wide Web (Schäfer, 2015). The corpus is split up into seven slices, each containing approximately 37 million sentences, but only the first slice was used in the current research project. The corpus considers punctuation marks as individual tokens, therefore separating those from the word they precede and follow. This treatment of punctuation marks is, however, not always correct in Dutch; apostrophes can be part of a word, for instance before a plural suffix (e.g. *taxi's* 'cabs'), which means that these parts of the word (i.e. the word *taxi*, the apostrophe and the plural suffix) needed to be reattached in the corpus. Frank and Hoeks (2019) preprocessed the corpus, reattaching the word parts of the words with an apostrophe, and the first preprocessed corpus slice was used in the current training process.

The final training data set was constructed in the following way. I started by creating the vocabulary. First, the 20,000 most frequent words of the corpus slice were selected for the vocabulary, excluding words with a non-letter (other than the hyphen or the apostrophe). Second, the word types from the set of test sentences that not yet appeared in the vocabulary were added. As a result, the final vocabulary comprised 20,153 word types in total. Subsequently, sentences were selected from the corpus slice that contained only words that occurred in the previously formed vocabulary. From that set of sentences, only those were kept in the training data set that (1) contained at least two words, (2) were not longer than 50 words (including punctuation tokens), and (3) contained no other punctuation token than the period, the comma, the exclamation point or the question mark. Consequently, the final training data set comprised 12,004,362 sentences (159,550,592 tokens).

6.2.2 Architecture details and training process

One LSTM network was trained on the training data set for two epochs. This means that the network went through the training data set twice. The goal of the training process was for the network to learn to predict the next word in a sequence. First during the training process, the words in the vocabulary go through a 300-unit word embedding layer, in which the words are transformed to a vector representation. The network learns where to position this vector, representing a word, in a continuous vector space from the context surrounding the word in the input sentences during the training process, and this position is called the word's embedding (Brownlee, 2021). These word vectors are then passed to a 600-unit recurrent layer (i.e. the LSTM block) and a 300-unit non-recurrent layer to optimize the network's performance. Last, the vectors are passed to the softmax output layer, which receives the output vectors of the recurrent layer as input, and converts these to a so-called next-word probability distribution for each word, which is the probability that the word will be the next word in a sequence (Crivellari & Beinat, 2020).

During training, gradients were clipped at 0.25. The network learns during training by updating the weights of the vector representations. These updates can be too large, however, thereby causing 'exploding gradients' (i.e. the weights over- or underflow). Gradient clipping is used to overcome this problem by setting a minimum and maximum value for the gradient for when it explodes (Goodfellow et al., 2016).

6.3 Evaluation process

After completion of the training process, the same experimental and control items that were judged by the human native speakers (discussed in Section 5.1.2) were presented as test data to the LSTM network. To evaluate the neural network's performance, the surprisal values were collected that the network assigned to the words in the test sentences. The surprisal metric comes from the surprisal theory, first proposed by Hale (2001), and is calculated as "the negative log probability of a word (w_i), given its preceding context($w_1 \ldots w_{i-1}$)" (Lowder et al., 2018, p.3):

$$S(w_i) = -log\ P(w_i|w_1 \dots w_{i-1})$$

A word's surprisal thus shows to what extent the word is unexpected in its context, with higher surprisal values representing higher unexpectedness and lower surprisal values representing lower unexpectedness. For the experimental items, for both non-islands and islands, surprisal was measured (1) at the verb immediately following the (filled) gap, e.g. *maakt* 'makes' for *wh*-islands and *aanbiedt* 'offers' for coordinate structure islands, and (2) summed over all words immediately following the (filled) gap, e.g. *(...) maakt in de bakkerij*

‘(…) makes in the bakery’ for *wh*-islands and *(...) aanbiedt tijdens de veiling* ‘(…) offers during the auction’ for coordinate structure islands (see Table 6 and Table 7 in Section 5.1). These specific regions are studied, because whether a gap is licit in the sentence or not should affect the network’s expectation locally and globally (Wilcox et al., 2018). For the control items, following Wilcox et al. (2019b), surprisal was measured summed over the entire sentence, and additionally it was measured at the critical possessive pronoun *zijn* ‘his’ or *haar* ‘her’.

6.4 Data analysis

Similar to the data analysis of the experimental part, all independent variables were coded using simple contrasts. The same coding scheme for the simple contrasts was used, which can be found in Table 9.

Four LME models were fitted, two for the experimental items and two for the control items. First, for the experimental items, I fitted one LME model with the single-word surprisal and one with the summed surprisal as the outcome variable, and both with PRESENCE OF GAP, PRESENCE OF FILLER, STRUCTURE, ISLAND and their interaction as fixed factors, using the *lmer* function from the *lmerTest* package (Kuznetsova et al., 2017) in R (version 3.6.0; R Core Team, 2019). Before fitting the model with single-word surprisal as the outcome variable, a box-cox transformation was performed on the surprisal values (with $\lambda = 1.52$) so that the transformed data was as close to normally distributed as possible. This was not necessary for the the model with summed surprisal as outcome variable. Second, for the control items, one LME model was fitted to the summed surprisal and one to the single-word surprisal with GENDER MATCH, STRUCTURE and their interaction as fixed effects. Before fitting the models, a logarithmic transformation was performed on the surprisal values as the box-cox indicated that this was necessary ($\lambda_{\text{summed}} = 0.10$; $\lambda_{\text{single-word}} = -0.14$). The random effects for all three models were based on the minimal AIC. After fitting the models, model assumptions were checked, which showed that the fitted values and the size of the error (ε) correlated for the two summed surprisal models (see Appendix C). Significance values for the coefficients from the two models were calculated using the Satterthwaite approximation in *lmerTest* (Kuznetsova et al., 2017). The interaction effects and three-level main effect were further examined using contrasts from the *emmeans* package (Lenth, 2022) in R.

6.5 Results

6.5.1 Analysis of the experimental items

Table 16 shows the mean single-word surprisal values per condition, and Table 17 the mean summed surprisal values per condition. Both are also illustrated in Figure 12.

Table 16

Mean single-word surprisal values. Associated standard deviations are given between parentheses.

		Non-island		Island	
Gap?	Filler?	*Wh*-island	Coordinate structure island	*Wh*-island	Coordinate structure island
Yes	Yes	12.20 (2.89)	11.50 (3.72)	12.60 (2.37)	11.90 (3.34)
Yes	No	13.00 (2.14)	12.50 (3.34)	13.10 (1.91)	12.10 (3.27)
No	Yes	11.60 (3.81)	10.60 (4.76)	11.60 (3.36)	10.30 (4.48)
No	No	11.80 (3.68)	10.80 (4.61)	11.80 (3.29)	10.40 (4.47)

Table 17

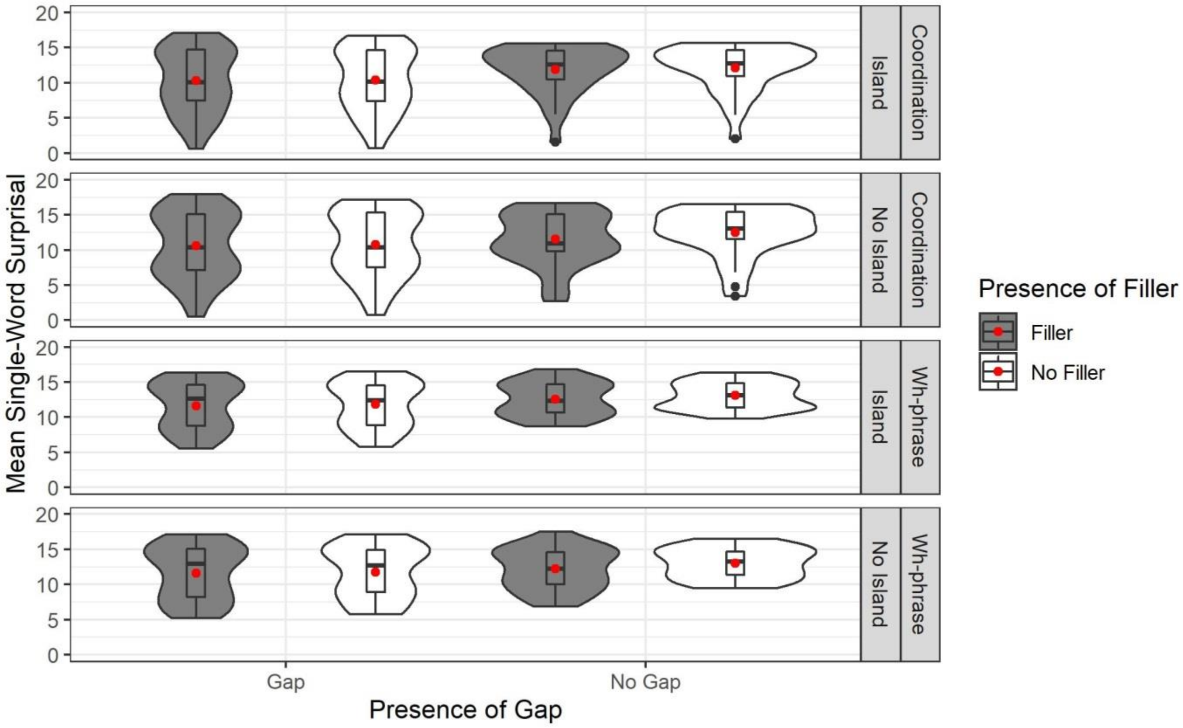
Mean summed surprisal values. Associated standard deviations are given between parentheses.

		Non-island		Island	
Gap?	Filler?	*Wh*-island	Coordinate structure island	*Wh*-island	Coordinate structure island
Yes	Yes	29.80 (5.94)	30.60 (6.95)	29.90 (5.18)	31.90 (6.60)
Yes	No	31.10 (5.63)	31.90 (6.53)	31.10 (5.15)	33.00 (6.56)
No	Yes	28.50 (6.24)	29.10 (7.34)	28.60 (5.84)	28.50 (7.05)
No	No	28.60 (6.30)	29.30 (7.42)	28.80 (6.02)	28.80 (7.28)

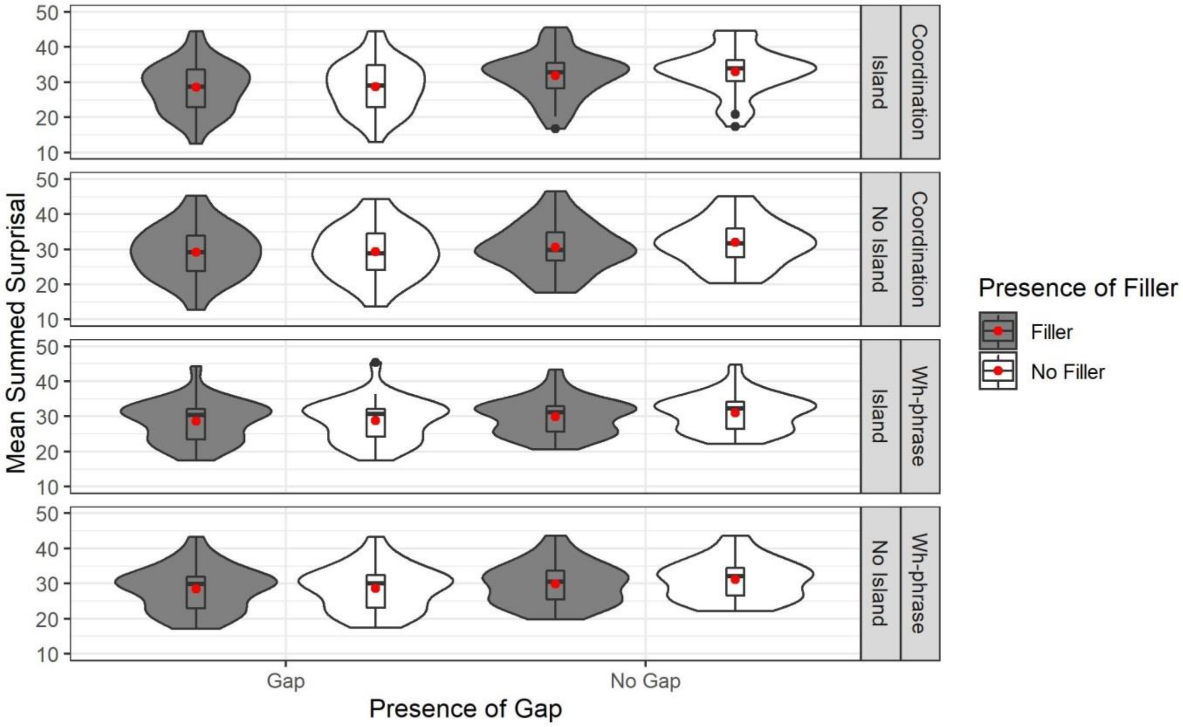
Figure 12

Violin/boxplot with the single-word (A) and summed (B) surprisal values on the y-axis, the levels of PRESENCE OF GAP *(gap vs. no gap) on the x-axis, the levels of* PRESENCE OF FILLER *(filler vs. no filler) representing the different colors, and the levels of* STRUCTURE *and* ISLAND *representing the different boxes. The red dot represents the mean.*

A.



B.



The final model included a random intercept and random slope for ISLAND and PRESENCE OF GAP for items. The results of the LME regression analyses are summarized in Table 18.

Table 18

Results of the LME regression analyses of the experimental items.

Fixed effects						
Predictors	Outcome variable	β [95% CI]	SE	df	*t*	*p*
Structure	Single-word surprisal	2.69 [−2.95, 8.32]	2.87	489	.94	.353
	Summed surprisal	−.84 [−3.90, 2.23]	1.56	489	−.54	.592
Island	Single-word surprisal	−.26 [−.89, .38]	.32	489	−.80	.429
	Summed surprisal	.21 [−.13, .54]	.17	489	1.21	.232
Presence of Filler	Single-word surprisal	−1.30 [−1.63, −.98]	.17	489	−7.89	< .001
	Summed surprisal	−.70 [−.85, −.55]	.08	489	−9.33	< .001
Presence of Gap	Single-word surprisal	3.71 [2.41, 5.01]	.66	489	5.59	< .001
	Summed surprisal	2.39 [1.83, 2.94]	.28	489	8.45	< .001
Structure x Presence of Gap	Single-word surprisal	−1.17 [−3.77, 1.44]	1.33	489	−.88	.383
	Summed surprisal	−1.12 [−2.23, −.00]	.57	489	−1.98	.053
Island x Presence of Filler	Single-word surprisal	.90 [.25, 1.55]	.33	489	2.72	.007
	Summed surprisal	.04 [−.26, .33]	.15	489	.26	.793
Island x Presence of Gap	Single-word surprisal	.85 [.20, 1.50]	.33	489	2.57	.011
	Summed surprisal	.82 [.53, 1.12]	.15	489	5.49	< .001
Presence of Filler x Presence of Gap	Single-word surprisal	−1.76 [−2.41, −1.11]	.33	489	−5.33	< .001
	Summed surprisal	−1.04 [−1.33, −.74]	.15	489	−6.91	< .001
Structure x Island x Presence of Gap	Single-word surprisal	−.35 [−1.65, .94]	.66	489	−.54	.592
	Summed surprisal	−1.91 [−2.50, −1.32]	.30	489	−6.34	< .001
Island x Presence of Filler x Presence of Gap	Single-word surprisal	1.79 [.50, 3.10]	.66	489	2.72	.007
	Summed surprisal	.25 [−.34, .84]	.30	489	.83	.408

Random effects			
Group	Name	Variance	SD
Item	Intercept		
	Single-word	131.24	11.46
	Summed	38.86	6.23
	Island		
	Single-word	4.97	2.23
	Summed	1.53	1.24
	Presence of Gap		
	Single-word	26.44	5.14
	Summed	4.75	2.18
Residual	Single-word	3.49	1.87
	Summed	.72	.85

Note. Single-word model: marginal R^2 = .05, conditional R^2 = .98; Summed model: marginal R^2 = .04, conditional R^2 = .98; only significant interaction effects with STRUCTURE included in this table.

The analysis first revealed significant main effects of PRESENCE OF FILLER (single-word: β = −1.30, SE_β = .17, 95% CI of β [−1.63, −.98], p < .001; summed: β = −.70, SE_β = .08, 95% CI of β [−.85, −.55], p < .001) and PRESENCE OF GAP (single-word: β = 3.71, SE_β = .66, 95% CI of β [2.41, 5.01], p < .001; summed: β = 2.39, SE_β = .28, 95% CI of β [1.83, 2.94], p < .001), but not of STRUCTURE (single-word: β = 2.69, SE_β = 2.87, 95% CI of β [−2.95, 8.32]; summed: β = −.84, SE_β = 1.56, 95% CI of β [−3.90, 2.23]) and ISLAND (single-word: β = −.26, SE_β = .32, 95% CI of β [−.89, .38]; summed: β = −.21, SE_β = .17, 95% CI of β [−.13, .54]). This shows that sentences without fillers ($M_{single\text{-}word}$ = 11.90, $SD_{single\text{-}word}$ = 3.53; M_{summed} = 30.30, SD_{summed} = 6.51) and sentences without gaps ($M_{single\text{-}word}$ = 11.10, $SD_{single\text{-}word}$ = 4.08; M_{summed} = 28.80, SD_{summed} = 6.63) were less unexpected than sentences with fillers ($M_{single\text{-}word}$ = 11.50, $SD_{single\text{-}word}$ = 3.68; M_{summed} = 29.60, SD_{summed} = 6.44) and gaps ($M_{single\text{-}word}$ = 12.40, $SD_{single\text{-}word}$ = 2.94; M_{summed} = 31.20, SD_{summed} = 6.10) respectively, but that there was no significant surprisal difference between non-islands ($M_{single\text{-}word}$ = 11.80, $SD_{single\text{-}word}$ = 3.74; M_{summed} = 29.90, SD_{summed} = 6.58) and islands ($M_{single\text{-}word}$ = 11.70, $SD_{single\text{-}word}$ = 3.49; M_{summed} = 30.10, SD_{summed} = 6.38) nor between the different island types *wh*-islands ($M_{single\text{-}word}$ = 12.20, $SD_{single\text{-}word}$ = 3.02; M_{summed} = 29.50, SD_{summed} = 5.81) and coordinate structure islands ($M_{single\text{-}word}$ = 11.30, $SD_{single\text{-}word}$ = 4.07; M_{summed} = 30.04, SD_{summed} = 7.07).

In addition, the analysis revealed a three-way interaction effect between ISLAND, PRESENCE OF FILLER and PRESENCE OF GAP (β = 1.79, SE_β = .66, 95% CI of β [.50, 3.10], p = .007), but only for the single-word surprisal model. Both for non-islands and islands, when there is a gap present in the sentence, the presence of a filler is more expected than its absence (non-islands: β = 3.08, SE_β = .33, p < .001; islands: β = 1.29, SE_β = .33, p = .003). On the other hand, also similar for non-islands and islands, the presence of a filler does not affect the surprisal values of the sentences without a gap (non-islands: β = .42, SE_β = .33, p = .907; islands: β = .42, SE_β = .33, p = .905). The three-way interaction is then caused by the different effect size of this surprisal pattern within non-islands and islands; the β-value indicates that the effect of the presence of a filler on the surprisal values in sentences with a gap is lower for islands (see Table 16 and Table 17 for means and standard deviations).

6.5.2 Analysis of the control items

Table 19 shows the mean single-word surprisal values per condition and Table 20 the mean summed surprisal values, which are both also illustrated in Figure 13.

Table 19

Mean single-word surprisal values. Associated standard deviations are given between parentheses.

Match?	No island	*Wh*-island	Coordinate structure island
Yes	5.33 (.76)	5.49 (1.15)	5.35 (.75)
No	5.44 (.71)	5.52 (1.06)	5.46 (.71)

Table 20

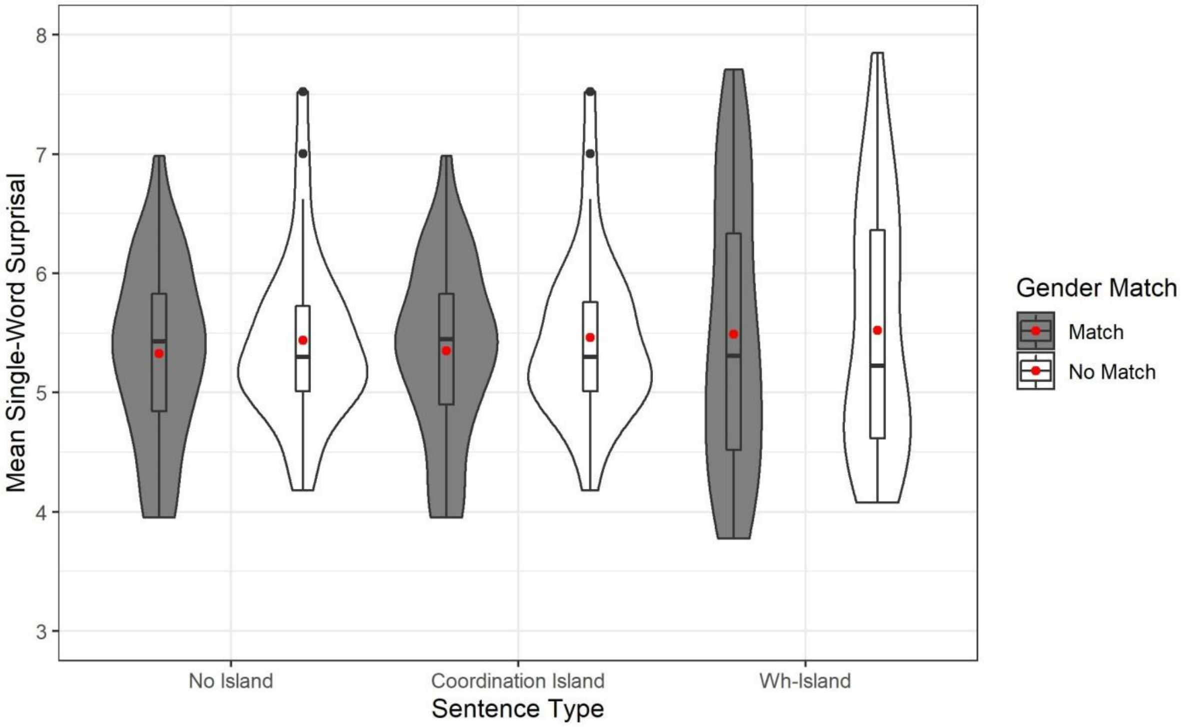
Mean summed surprisal values. Associated standard deviations are given between parentheses.

Match?	No island	*Wh*-island	Coordinate structure island
Yes	78.80 (8.49)	82.70 (8.85)	90.70 (8.76)
No	78.90 (8.50)	82.60 (8.83)	90.70 (9.02)

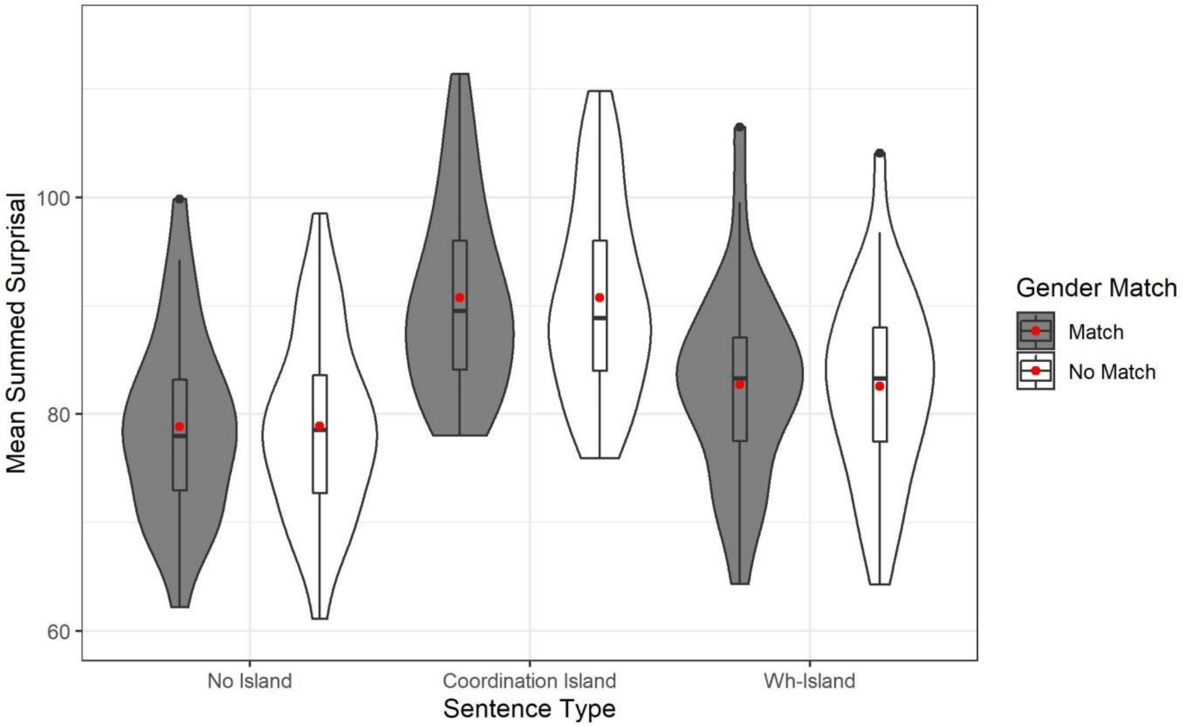
Figure 13

Violin/boxplot with the single-word (A) and summed (B) surprisal values on the y-axis, the levels of ISLAND *(non-island vs. wh-island vs. coordinate structure island) on the x-axis, the levels of* GENDER MATCH *(match vs. no match) representing the different colors. The red dot represents the mean.*

A.



B.



The final single-word surprisal model included a random intercept and random slope for GENDER MATCH for items, and the summed surprisal model for ISLAND and GENDER MATCH for items. The results of the LME regression analyses are summarized in Table 21.

Table 21

Results of the regression analysis of the control items.

Fixed effects						
Predictors		β [95% CI]	SE	df	*t*	*p*
Island 1 (no island vs. coordinate structure island)	Single-word	.00 [−.02, .03]	.01	182	.34	.736
	Summed	.14 [.13, .15]	.01	175	27.40	< .001
Island 2 (no island vs. *wh*-island)	Single-word	.01 [−.01, .04]	.01	182	.89	.373
	Summed	.05 [.03, .06]	.01	175	7.16	< .001
Match	Single-word	.02 [−.06, .09]	.04	182	.48	.637
	Summed	−.00 [−.01, .01]	.00	175	−.17	.867
Match x island 1	Single-word	−.00 [−.05, .05]	.03	182	−.01	.993
	Summed	−.00 [−.01, .00]	.00	175	−.34	.732
Match x island 2	Single-word	−.01 [−.07, .04]	.03	182	−.49	.626
	Summed	−.00 [−.01, .00]	.00	175	−1.04	.300

Random effects			
Group	Name	Variance	SD
Item	Intercept		
	Single-word	.01	.09
	Summed	.01	.10
	Island 1		
	Single-word	.00	.03
	Island 2		
	Single-word	.00	.04
	Match		
	Single-word	.04	.21
	Summed	.00	.02
Residual	Single-word	.01	.08
	Summed	.00	.01

Note. Single-word model: marginal R^2 = .00, conditional R^2 = .77; Summed model: marginal R^2 = .24, conditional R^2 = .99.

Neither analysis revealed a main effect of GENDER MATCH (single-word: β = .02, SE_β = .04, 95% CI of β [−.06, .09]; summed: β = −.00, SE_β = .01, 95% CI of β [−.01, .01]), nor an interaction between ISLAND and GENDER MATCH (single-word: F(2, 124) = .16, p = .856; summed: F(2, 157.04) = .01, p = .986). This means that no gendered expectation effect was found, represented by a non-significant difference between sentences with and without a gender match, not for any of the ISLAND levels.

6.6 Discussion

To investigate to what extent an artificial neural network is sensitive to the *wh*- and coordinate structure island constraints in Dutch, an LSTM network was tested on the same set of test sentences judged by the human native speakers in the previous section. To assess its sensitivity, the surprisal values it assigned to the words in these test sentences were examined, representing how unexpected a word was in a sequence. Equal to the acceptability judgement task, the experimental design was based on the interaction design introduced by Wilcox et al. (2018), and a set of control items was added as suggested by Wilcox et al. (2019b).

While it was predicted that sentences with a gender match would be less surprising than sentences with a mismatch, the LSTM network did not show this gendered expectation effect; the surprisal values do show a numerical trend towards this difference, but the mean surprisal values for gender match and mismatch sentences do not significantly differ in the current experiment. Moreover, no interaction effect was found, suggesting that neither non-islands nor islands showed a gendered expectation effect. While this seems to show that the network is not able to thread information through islands, the absence of a gendered expectation effect in non-islands suggests that the LSTM network is not able to maintain expectancies at all, at least for gender. There are several possible explanations for the absence of the gendered expectation effect.

First, it could be the case that one of the two possessive pronouns is more frequent in the training data set than the other. This could have affected the network's performance as it might have only been able to accurately learn the gendered expectation effect for one of the genders. While this explanation should have been ruled out by the counterbalancing of the genders within each level of GENDER MATCH, I tested for this alternative explanation. When performing the analyses described in Section 6.5.2 for the male items only or for the female items only, still no gendered expectation effect was found (p_{male} = .932; p_{female} = .524). The absence of the effect can thus not be explained by the frequency of the possessive pronouns in the training data set. Second, it could also be the case that the full NPs with un unambiguous gender used in the items might not be frequent enough in the training data set for the neural network to learn to associate it with a clear gender (see Appendix D for frequencies per full NP used). Last, previous literature suggests that LSTM networks have more difficulty with dependencies when the sentence has more embedding layers (Chowdhury & Zamparelli, 2018). As the current test items all had three layers of embedding, similar to the English items in Wilcox et al. (2021), this could have caused the difficulty in maintaining a gender expectancy through these layers in Dutch. In a future experiment, the frequency of the full NPs and the number of embedding layers should therefore be controlled for.

All in all, it is clear that the control experiment implemented to control for a complexity effect is inconclusive in the current experiment. This means that the results discussed below should be interpreted with caution, as they could possibly be an effect of complexity.

When looking at the results of the experimental items, it is again important to note that the interaction design by Wilcox et al. (2018) assumes that (1) gaps require fillers and (2) fillers require gaps. For non-islands, if the network learned the first assumption, gaps should be more surprising when no *wh*-filler is present. The results show that this assumption is learned; the absence of a *wh*-filler increases the surprisal on the post-gap verb in sentences with a gap.

Moreover, if the network also learned the second assumption, filled argument positions should be more surprising when a *wh*-filler is present. Contrary to the first assumption, however, this second assumption does not seem to be learned by the network; when processing a sentence without a gap, the presence of a *wh*-filler did not affect the surprisal on the post-filled-gap verb. This means that even when the network crossed a *wh*-filler, it did not create an expectation for a gap further on in the sentence, and that the ungrammatical sentence with no gap but a filler is as (un)expected as the perfectly grammatical sentence with neither a filler nor a gap. While not creating an expectation for a gap is similar to what was observed in the control items for gender, it is remarkable that a gap expectation was made when there was an actual gap present later in the sentence. This all suggest that having an argument too little might be worse for the network than having an argument too many in the sentence. That is, coming across a filler and not having a gap to link that filler to, i.e. having an argument too many, does not seem to matter for the network in comparison to when it has not seen a filler but comes across a gap that needs one, i.e. having an argument too little.

Also important to note is that this result only showed up when the surprisal was measured at the immediate post-(filled-)gap verb, and not when it was summed across all post-(filled-) gap material. This shows that having no gap to link the filler to did also not increase the surprisal later in the sentence. The word or region where surprisal was measured might, however, also explain the result found for non-islands. In the current research project, regardless of measuring only at the verb or over the whole post-filled-gap region, I always started measuring surprisal after the filled gap. The surprisal that is predicted to occur, however, might only be observable at the filled gap itself. Instead of looking at the full *wh*-licensing interaction, as was done in the current study, Wilcox et al. (2019a) investigated the *wh*-effects in the +Gap and –Gap conditions individually. In this way, they could vary the word/region that surprisal was measured at between these conditions; in the +Gap condition, the surprisal was measured at the verb immediately following the gap, while in the –Gap condition, the surprisal was measured at the filled-gap. For future research, it would be interesting to find out whether taking on Wilcox et al.'s (2019a) approach affects the current results.

Next, for islands, it was also hypothesized that filled argument positions should be more surprising when a *wh*-filler is present. Similar to what was observed for non-islands, however, the presence of a filler did not affect the surprisal measured at the immediate post-filled-gap verb. Again, this suggests that coming across a filler and not having a gap to link that filler to, i.e. having an argument too many, does not seem to affect the network. The measurement site could not be a possible explanation for this effect, however; as the network should never expect a gap inside an island, a filled-gap position within an island should not surprise the network.

When gaps appear inside island configurations, however, hypotheses were different; the presence of *wh*-filler should not affect the network's expectations in island configurations containing a gap, as these sentences are always ungrammatical. This hypothesis is not supported by the current results, however. When islands contained a gap, the presence of a filler decreased surprisal values. This shows that the LSTM network treats islands no different from non-islands; when there is a gap, there must be a filler. The network, trained on 12 million Dutch sentences, was thus not able to recognize island configurations and block the gap expectancies within these structures. While a three-way interaction was thus found, this seems to be solely driven by the difference in effect size as opposed to effect direction.

Last, although not part of Wilcox et al.’s (2018) interaction design, I also predicted a difference between the two island types; the network should be more surprised to see a coordinate structure island in comparison to a *wh*-island, and the three-way interaction described above should be more pronounced for coordinate structure islands as well. This prediction was not borne out, as there was no difference in surprisal between the two island types. The results found in the current experiment suggest that neither a coordinate structure island effect nor a *wh*-island effect was found for the LSTM network in Dutch.

Now that the results of the acceptability judgement and the modelling task have been discussed individually, I will compare these results and discuss their implications for the debate about language acquisition in the next section.

7. General discussion

In this section, I will start by summarizing the current research project. Next, as I have already separately discussed and interpreted the results of the native speakers and the LSTM network, I will compare these to see whether the LSTM network's sensitivity to the *wh*- and coordinate structure island constraints is comparable to that of humans, and I will discuss what these results mean for the debate about language acquisition. Last, I will discuss the strengths and limitations of the current research project and tie these to suggestions for future research.

7.1 Summary of the current research project

The current research project investigated whether artificial neural networks show the same sensitivity to *wh*- and coordinate structure island violations as human native speakers do in Dutch. First, it was established whether these constraints exist in Dutch, and if so, to what extent native speakers are sensitive to them. Next, an LSTM network was tested on its sensitivity to these constraints using the same materials and experimental design.

To test the native speakers, an acceptability judgement task was performed using the interaction design introduced by Wilcox et al. (2018). For each island type, the experimental items were manipulated for PRESENCE OF GAP (no gap vs. gap), PRESENCE OF FILLER (no filler vs. filler) and ISLAND (non-island vs. island). Example sentences of all these conditions can be found in Table 6 and Table 7. The native Dutch speakers had to judge on a 7-point scale whether the regular filler-gap dependencies, *wh*-islands and coordinate structure islands with and without gaps and/or fillers were acceptable in Dutch.

The results showed that, for regular filler-gap dependencies without any island structures, humans correctly judged that gaps require fillers and that fillers require gaps. Additionally, when presented by island configurations, native Dutch speakers showed an equally large *wh*-island and coordinate structure island effect by showing that island configurations were only acceptable without gaps and fillers. Moreover, a control study was used to rule out the possibility that the native speakers only showed these island effects because of the island's complex syntactic structure. The current results thus suggest that the *wh*- and coordinate structure island constraints exist in Dutch as the native speakers showed sensitivity to them.

After the acceptability judgement task, an LSTM network, trained on 12 million sentences extracted from the Dutch *Corpora Of the Web* (NLCOW14), was presented with the same test sentences. To assess the network's sensitivity to the island constraints, the surprisal values it assigned to each word in the test sentences were collected, which indicated the extent to which a word was unexpected by the network.

The results showed that, when processing regular filler-gap dependencies, the network could correctly recognize that gaps are more expected when there is a *wh*-filler present in the sentence. On the other hand, it did not learn that a *wh*-filler should be more unexpected when there is no gap in the sentence, but a filled argument position. Remarkably, however, within island configurations, exactly the same pattern was found. This means that the network does not treat islands different from non-islands; when there is a filler, there must be a gap, even when that gap occurs inside an island. An LSTM network, trained on 12 million Dutch sentences, does thus not seem able to recognize island configurations in Dutch. There are several reasons possible to explain why, as the control study in the modelling experiment was inconclusive. The network's results could thus be explained by the complexity of the sentences used, by the architectural details of the network, by the structural properties of Dutch, or by the absence of innate language knowledge or abilities in the network. These reasons will be discussed in more detail below.

7.2 Comparing Dutch native speakers and the LSTM network

To assess whether the sensitivity to *wh*- and coordinate structure islands of the LSTM network are actually similar to what human native speakers of Dutch show, it will first be examined whether the same statistical effects were found in both the judgement and the modelling task. The results of these tasks showed that both in the acceptability judgement task and the modelling task, a three-way interaction effect was found between PRESENCE OF GAP, PRESENCE OF FILLER and ISLAND, as hypothesized in Section 4. While this seems to indicate that the LSTM network thus performed similar to the native speakers, it is crucial to investigate this effect in more detail. In this investigation, the acceptability ratings and the surprisal values will be compared using the translation proposed by Pearl and Sprouse (2015); the higher the surprisal value, the lower the acceptability rating, and the lower the surprisal value, the higher the acceptability rating. The values relevant to explain the three-way interaction effect in both tasks are summarized side by side in Table 22.

Table 22

Summarized standardized acceptability judgements, single-word surprisal values and statistical significance of differences per experimental condition. Standard deviations are given between parantheses.

		Non-island				Island			
Gap?	Filler?	Judgements (stan.)		Surprisal (single-word)		Judgements (stan.)		Surprisal (single-word)	
Yes	Yes	−.46 (.70)	**	11.90 (3.32)	***	−.82 (.48)	***	12.20 (2.89)	**
Yes	No	−.59 (.66)		12.80 (2.79)		−.60 (.71)		12.60 (2.70)	
No	Yes	−.45 (.69)	***	11.10 (4.31)	.	−.48 (.67)	***	11.00 (3.99)	.
No	No	.63 (.73)		11.30 (4.17)		.65 (.73)		11.10 (3.96)	

Note. $^{**}p < .01$; $^{***}p < .001$.

I will compare the acceptability judgements and the surprisal values per assumption made by Wilcox et al.'s (2018) interaction design. Starting with non-islands, both native Dutch speakers and the LSTM network assume that gaps require fillers; gaps are less acceptable and more surprising when no *wh*-filler is present in the sentence. In contrast to this first assumption, humans and the network do not align on whether they have mastered the second assumption, namely that fillers require gaps. While the presence of a *wh*-filler did decrease acceptability of sentences without a gap, it did not affect the surprisal on the post-filled-gap verb in these sentences. This means that humans seem to have learned the second assumption, but the LSTM network did not; even when the network crossed a *wh*-filler, it did not create an expectation for a gap further on in the sentence, meaning that the ungrammatical sentence with no gap but a filler is as (un)expected as the perfectly grammatical sentence with neither a gap nor a filler. As discussed in Section 6.6, this could be due to the measurement site used;

the surprisal that is predicted to occur might only be observable at the filled gap itself (Wilcox et al., 2019a).

Moving on to island configurations, it was predicted that sentences with a gap inside the island should never be acceptable and always surprising, regardless of having a *wh*-filler present in the sentence. That is because having a gap inside an island is always ungrammatical in Dutch. Native speakers indeed judged all sentences with gaps in the ungrammatical range of the 7-point scale. The presence of a *wh*-filler, however, decreased acceptability of these sentences even more. While it was not predicted per se, this result is perfectly explainable; humans create an expectation for a gap somewhere in the sentence when they cross the *wh*-filler at the start of the sentence, but this expectation is never fulfilled as there is no licit gap position available in the sentence. Interestingly, the LSTM network showed exactly the opposite behaviour; the presence of a *wh*-filler decreased surprisal values of sentences with a gap inside an island. While humans thus found islands with a gap more unacceptable when they encountered a *wh*-filler, it made them less unexpected for the network. This shows that the network treats these gaps as licit; it does not recognize *wh*- and coordinate structure islands and does not block gap expectancies within these structures, unlike the human native speakers of Dutch.

To further substantiate this last claim, the speakers and the network also diverge in their behaviour on sentences without gaps. While humans expectedly judge sentences with filled argument positions as less acceptable when a *wh*-filler is present, the presence of a filler did not affect the surprisal assigned to the post-filled-gap verb by the network. Again, the network treats island configurations no different from non-island configurations, and clearly differs with this from the human native speakers of Dutch.

In sum, after comparing the acceptability judgements by humans and the surprisal values by the LSTM network, it can only be concluded that, unlike humans, the LSTM network is not able to recognize *wh*- and coordinate structure islands and to block gap expectancies in these islands. This means that, while humans show a strong sensitivity to the *wh*-island and coordinate structure island constraints, the LSTM network does not in Dutch. Possible reasons as to why this is the case, and the implications of this result for the debate on language acquisition will be discussed in the next section.

7.3 Implications

While the human native speakers of Dutch seem to show a sensitivity to *wh*- and coordinate structure island violations, the LSTM network does not seem able to learn to recognize these gap-resistant structures. As already discussed, this could be due to the experimental design and/or analysis of the current research project, the network's architecture, the structural properties of Dutch, or the absence of innate language abilities in the network. The latter three reasons will be discussed below.

First of all, the results could be explained by the fact that the network simply did not learn Dutch well enough, or at least not as well as the LSTM networks seemed to have learned English in Wilcox et al. (2021). While not much is said about the specific architectural details of the networks used in Wilcox et al. (2021), the architectural difference between their networks and the current network could have led to the difference in learning success. Future research should investigate this explanation in detail by comparing the architecture of the networks used in both studies. One thing that can already be compared between the two studies is the size of the training data set; Wilcox et al. (2021) trained one LSTM network on 90 million tokens and the other on roughly 1 billion tokens, while the current training data set consisted of approximately 160 million tokens. Their results, however, do not show any qualitative differences between the learning success of their two networks. If the network trained on their smallest training data set performed no different from the network trained on their largest data set, there is no reason to believe that the size of the current data set affected the network's performance as compared to Wilcox et al. (2021). While the quantity of the training data should thus not be of influence, the quantity of this data set could have an effect. This will be discussed more below.

Second, as discussed in Section 3.2, recent literature suggests that artificial neural networks have a performance bias for right-branching structures, and thus also for right-branching languages like English (Dyer et al., 2019). While this right-branching bias will thus inflate the architecture's performance in right-branching languages, it will undermine its performance in left-branching and possibly mixed-branching languages (Li et al., 2020). As Dutch has mixed-branching, the network's inability to learn the island constraints could be tied to the fact that Dutch is not fully right-branching. Therefore, it could be interesting for future research to investigate whether networks can learn about this mixed branching in Dutch, for example by examining sentence embedding. As discussed above, this is relevant to see whether it can handle the complexity of having multiple embedding layers, but it can also contribute to the branching direction explanation. When a Dutch sentence contains a matrix and an embedded sentence, two different branching directions are used; the embedded sentence uses the basic word order SOV (Subject-Object-Verb), but the matrix sentence uses the word order SVO (Subject-Verb-Object) due to movement of the verb. As the current research project only used embedded sentences, and thus a combination of both branching directions, it will be interesting to see in future research whether this could have affected the network's performance. Furthermore, to directly test for the influence of branching direction, languages with different branching directions can be compared directly.

Third, while the training data set should have been big enough to be able to learn, the information in this data might not have been enough for the neural network to learn about the syntactic island constraints, as many syntacticians have suggested before (Chomsky, 1965; Pearl & Sprouse, 2013). While children can already recognize syntactic islands at the age of four (Bates & Pearl, 2021), the network does not seem able to do the same with current training data covering a lot more than four years of a human's life (Wilcox et al., 2021). This could suggest that children use something else than just external input to learn the syntactic island constraints. While the nature of this "something else" is still disputed, it is internal, such as internal language knowledge or abilities. This shows that the result of the current research project could add relevant new insights into the debate about language acquisition.

However, it must be noted that the syntactic learning inability of a neural network could never be able to provide direct evidence about an innate language ability as there are still too many differences between the architecture of the artificial neural network and the human brain. The failure of the network in the current research study to learn the syntactic island constraints does thus not mean that syntactic island constraints cannot be learned. It does, however, suggest that a domain-general learner is, at least in the current study, not able to recognize island configurations. It was also shown that this domain-general behaved differently than humans, who have been argued to be domain-specific learners.

While this research could thus provide relevant new insights for the debate about language acquisition, it can also provide relevant new knowledge to the field of experimental syntax and shed new light on the debate about the source of island effects, which was discussed in Section 2.1. In the acceptability judgement task, it was found that native Dutch speakers judge *wh*- and coordinate structure island violations as unacceptable as compared to non-islands. This suggests that the *wh*- and coordinate structure island constraints exist in Dutch. This strengthens the results found by Beljon et al. (2021), namely that native Dutch speakers are sensitive to the *wh*-island constraint, and adds relevant new data to the field of experimental syntactic research on syntactic island constraints as Dutch is underrepresented here. Moreover, the current study is the first to experimentally confirm the theoretical claim that "it does not seem possible to extract one or more full conjuncts, in any language" (Liu et al., 2022, p. 503), at least for Dutch.

To tie into the debate about the source of these island effects found, the current research project employed a control study that suggested that the results found cannot be accounted for by an extra-syntactic complexity effect. Previous research suggested that native speakers might not be able to thread information through syntactically complex constructions like islands, causing the unacceptability of island violations (Keshev & Meltzer-Asscher, 2019). If processing indeed encounters difficulties in these constructions, however, native speakers should not have been able to maintain a gender expectancy within islands. The current results of the control study showed that they could, suggesting that the complexity explanation cannot account for the island effects found in the current research project. The results found by Beljon et al. (2021) also seem to suggest this for Dutch; using a complex *wh*-filler and adding a discourse context, which should decrease processing difficulty, only showed to have an effect in combination, with the effect being minimal. As an extra-syntactic explanation thus does not seem able to account for the results, it can be suggested that human native speakers of Dutch block gap expectancies in island configurations because of a syntactic constraint. This seems further supported by the fact that the LSTM network, a domain-general learner without any innate language abilities, was not able to learn the island constraints correctly. If this network would have been able to learn the island constraints comparable to the human speakers, it could not have used innate syntactic constraints. However, the learning inability of this network suggests that it does need something more than just input, possibly innate syntactic island constraints.

In sum, the results of the current research project provide relevant new insights for the debates about language acquisition and the source of island effects. However, it is important to note that much more research is necessary on this topic to be able to support the now still carefully formulated implications.

7.4 Strengths, limitations and future research

The current research project contained a strong theoretical background and methodological approach. However, some improvements would make it even more useful for future research in the same research area. The strengths and limitations of the current research project and suggestions for future research will now be discussed.

First, unlike many other studies performed before in this research field, the current research project subjected the human native speakers and the artificial neural network to the same scrutiny; the same experimental design, test sentences and control study were used to test both humans and a network. In this way, their results could be easily compared to assess whether the sensitivity to island constraints of the neural network was comparable to that of human speakers of Dutch (Futrell et al., 2018).

Second, in the creation of this experimental design, test sentences and control study, I closely followed the successful investigation on the learnability of syntactic islands by neural networks in English performed by Wilcox et al. (2021), combining insights from the studies by Wilcox et al. (2018) and Wilcox et al. (2019b). This, however, also led to some limitations that could help improve the methodological approach in the future.

The first limitation concerned the test items used. Following Wilcox et al. (2021), I created test sentences containing three embedding layers and a bare *wh*-filler. This made the sentences very complex, which might have caused the control study in the modelling experiment to be inconclusive and might have influenced the neural network's results. In future research, I would suggest to first investigate to what extent the ANN is affected by sentence complexity; e.g. if its performance is influenced by the number of embedding layers.

In addition, the second limitation concerned the data analysis. Following Wilcox et al. (2021), I compared the full *wh*-licensing interaction between non-islands and islands to see whether it reduced within island configurations. It is also possible, however, to statistically investigate the *wh*-effects in the +Gap and –Gap conditions separately, which together make up the licensing interaction. In this way, the region where surprisal is measured can also vary between these *wh*-effects (Wilcox et al., 2019a). As I have stated in Section 6.6, measuring surprisal on immediate post-gap material can be sufficient to see effects in the +Gap condition, but it might be better to check out the surprisal on the filled-gap in the –Gap conditions. This might alter the results found for the LSTM network in the current research.

Another strong point about the current research project is that it is theoretically relevant. As all research on the learnability of syntactic island constraints by neural networks has been performed in English, and neural networks seem to have a performance bias for English-like structural input, this research investigated whether an LSTM network can learn syntactic island constraints in Dutch. This was not only relevant for the research within computational psycholinguistics, but also provided relevant data and knowledge to the field of (experimental) syntax; Dutch is underrepresented in the experimental data about syntactic island constraints. This theoretical relevance creates many other opportunities to provide even more relevant insights to both these fields, which will be described below.

First, this research project showed that the LSTM network treated island configurations similar to non-island configurations. Therefore, it would be interesting to look at the latter in more detail. Future research should first investigate whether the neural network can learn all filler-gap characteristics in Dutch, as described in Wilcox et al. (2021) for English. If that is established, it will be interesting to include more island types, and other types than those used in the current research project. Not only will this provide relevant new data and knowledge for the field of experimental syntax, it can also show whether different island types can be learned by the neural network. In future studies, however, the control study should be improved; different control studies should be tried out to see which study is best to use in this kind of research in Dutch.

Second, when more research on this is (being) done in Dutch, it is also relevant to extend it to even more languages. As a first option, it would be interesting to extend it to a left-branching language to see whether branching direction indeed affects the network's performance.

In conclusion, the current research project showed that, while native speakers of Dutch show a strong sensitivity to *wh*- and coordinate structure island violations, an artificial neural network does not seem able to similarly recognize these island configurations and block gap expectancies within these structures. This suggests that input alone might not be enough to learn about syntactic island constraints, and that internal language knowledge or abilities might be necessary to learn about these constraints.

8. Abbreviations

MASC	masculine
REF	referential pronoun

9. Open data

The raw data and analysis scripts from this study can be found here: https://osf.io/kt3he/

10. Reference list

Abeillé, A., Hemforth, B., Winckel, E., & Gibson, E. (2020). Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition*, *204*(104293), 1–23. https://doi.org/10.1016/j.cognition.2020.104293

Bates, E. (2003). Natura e cultura nel linguaggio [On the nature and nurture of language]. In R. Levi-Montalcini, D. Baltimore, R. Dulbecco, F. Jacob, E. Bizzi, P. Calissano, & V. Volterra (Eds.), *Frontiere della biologia. Il cervello di Homo sapiens* (pp. 241–265). Istituto della Enciclopedia Italiana.

Bates, A., & Pearl, L. (2021). *When do input differences matter? Using developmental computational modelling to assess input quality for syntactic islands across socio-economic status* [Unpublished manuscript]. Department of Language Science, University of California.

Beljon, M., Joosen, D., Koeneman, O., Ploum, B., Sommer, N., De Swart, P., & Wilms, V. (2021). The effect of filler complexity and context on the acceptability of the wh-island violations in Dutch. In M. Dingemanse, E. Van Lier, & J. Vogels (Eds.), *Linguistics in the Netherlands* (Vol. 38, pp. 4–20). John Benjamins Publishing Company. https://doi.org/10.1075/avt.00047.bel

Brownlee, J. (2021, February 1). *How to Use Word Embedding Layers for Deep Learning with Keras*. Machine Learning Mastery. https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/

Chaves, R.P. (2020). What Don’t RNN Language Models Learn About Filler-Gap Dependencies? In *Proceedings of the Society for Computation in Linguistics* (pp. 20–30). Association for Computational Linguistics.

Chaves, R.P., & Putnam, M.T. (2020). *Unbounded Dependency Constructions: Theoretical and Experimental Perspectives*. Oxford University Press. https://doi.org/10.1093/oso/9780198784999.001.0001

Chomsky, N. (1964). *Current Issues in Linguistic Theory*. De Gruyter.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. (1971). *Problems of Knowledge and Freedom*. Fontana.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 232–286). Holt, Reinhart & Winston.

Chomsky, N. (1993). A minimalist program for linguistic theory. In K. Hale & S.J. Keyser (Eds.), *The View From Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. MIT Press.

Chopra, A., Prashar, A., & Sain, C. (2013). Natural language processing. *International Journal of Technology Enhancements and Emerging Engineering Research, 1*(4), 131–134. https://doi.org/10.1109/inmic.2004.1492945

Chowdhury, S.A., & Zamparelli, R. (2018). RNN Simulations of Grammaticality Judgments on Long-distance Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 133–144). Association for Computational Linguistics.

Christensen, K.R., Kizach, J., & Nyvad, A.M. (2013). Escape from the Island: Grammaticality and (Reduced) Acceptability of wh-island Violations in Danish. *Journal of Psycholinguistic Research*, *42*(1), 51–70. https://doi.org/10.1007/s10936-012-9210-x

Crivellari, A., & Beinat, E. (2020). LSTM-Based Deep Learning Model for Predicting Individual Mobility Traces of Short-Term Foreign Tourists. *Sustainability*, *12*(349), 1–18. https://doi.org/10.3390/su12010349

Da Costa, J.K., & Chaves, R.P. (2020). Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics* (Vol. 3, pp. 189–198). Association for Computational Linguistics. https://doi.org/10.7275/3sb6-4g20

Davis, F., & Van Schijndel, M. (2020). Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1979–1990). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.179

Dyer, C., Melis, G., & Blunsom, P. (2019). *A Critical Analysis of Biased Parsers in Unsupervised Parsing*. arXiv. https://doi.org/10.48550/arXiv.1909.09428

Elman, J.L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Frank, S. (2021). Cross-language structural priming in recurrent neural network language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, pp. 664–665). Cognitive Science Society.

Frank, S., & Hoeks, J.C.J. (2019). The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *Proceedings of the Cognitive Science Society* (pp. 337–343). Cognitive Science Society.

Frazier, L., & Rayner, K. (1988). Parameterizing the Language Processing System: Left- vs. Right-Branching within and across Languages. In J. A. Hawkins (Ed.), *Explaining Language Universals* (pp. 247–279). Basil Blackwell.

Frost, R., Armstrong, B.C., Siegelman, N., & Christiansen, M.H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125. https://doi.org/10.1016/j.tics.2014.12.010

Futrell, R., Wilcox, E., Morita, T., & Levy, R. (2018). *RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency*. arXiv. https://doi.org/10.48550/arXiv.1809.01329

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. In *Proceedings of NAACL-HLT 2019* (pp. 32–42). Association for Computational Linguistics.

Goldberg, A.E. (2014). Backgrounded constituents cannot be ‘extracted’. In J. Sprouse & N. Hornstein (Eds.), *Experimental Syntax and Island Effects* (pp. 221–238). Cambridge University Press.

Goldberg, Y. (2019). *Assessing BERT’s Syntactic Abilities*. arXiv. https://doi.org/10.48550/arXiv.1901.05287

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning series)* (Illustrated ed.). The MIT Press.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL-HLT 2018* (pp. 1195–1205). Association for Computational Linguistics.

Hale, J.T. (2001). A probabilistic early parser as a psycholinguistic model. In *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Association for Computational Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofmeister, P., & Sag, I.A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415. https://doi.org/10.1353/lan.0.0223

Hu, J. (2018, January 20). *A Simple Starter Guide to Build a Neural Network*. Towards Data Science. https://towardsdatascience.com/a-simple-starter-guide-to-build-a-neural-network-3c2cf07b8d7c

Kemmerer, D.L. (2015). *Cognitive Neuroscience of Language*. Amsterdam University Press.

Keshev, M., & Meltzer-Asscher, A. (2018). A processing-based account of subliminal wh-island effects. *Natural Language & Linguistic Theory*, *37*(2), 621–657. https://doi.org/10.1007/s11049-018-9416-1

Kovač, I., & Schoenmakers, G. (2022). An experimental-syntactic take on long passive in Dutch: Unraveling the patterns underlying its (non-)acceptability [Manuscript submitted for publication]. Department of Language and Communication, Radboud University.

Kush, D., Lohndal, T., & Sprouse, J. (2019). On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language*, *95*(3), 393–420. https://doi.org/10.1353/lan.2019.0051

Kuznetsova, A., Brockhoff, P., & Christensen, R., (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. https://doi.org/10.18637/jss.vo82.i13

Lappin, S., & Shieber, S.M. (2007). Machine learning theory and practice as a source of insightinto universal grammar. *Journal of Linguistics*, *43*(2), 393–427. https://doi.org/10.1017/s0022226707004628

Lau, J.H., Clark, A., & Lappin, S. (2016). Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, *41*(5), 1202–1241. https://doi.org/10.1111/cogs.12414

Lenth, R.V. (2022). *Emmeans: Estimated Marginal Means, aka Least-Squares Means* (version 1.7.2) [R package]. https://CRAN.R-project.org/package=emmeans

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

Li, H., Liu, L., Huang, G., & Shi, S. (2020). On the Branching Bias of Syntax Extracted from Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4473–4478). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.401

Linzen, T., & Baroni, M. (2021). Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. In *Transactions of the Association for Computational Linguistics* (Vol. 3, pp. 521-535).

Liu, Y., Winckel, E., Abeillé, A., Hemforth, B., & Gibson, E. (2022). Structural, Functional, and Processing Perspectives on Linguistic Island Effects. *Annual Reviews*, *8*, 495–525. https://doi.org/10.1146/annurev-linguistics-011619-030319

Lowder, M.W., Choi, W., Ferreira, F., & Henderson, J.M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive Science*, *42*, 1166–1183. https://doi.org/10.1111/cogs.12597

Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, L., & Linzen, T. (2020). Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5523–5539). Association for Computational Linguistics.

Newmeyer, F.J. (2016). Nonsyntactic Explanations of Island Constraints. *Annual Review of Linguistics*, *2*(1), 187–210. https://doi.org/10.1146/annurev-linguistics-011415-040707

Pañeda, C., Lago, S., Vares, E., Veríssimo, J., & Felser, C. (2020). Island effects in Spanish comprehension. *Glossa: a journal of general linguistics*, *5(1)*(21), 1–30. https://doi.org/10.5334/gjgl.1058

Pearl, L. (2021). Poverty of the Stimulus Without Tears. *Language Learning and Development*, 1–40. https://doi.org/10.1080/15475441.2021.1981908

Pearl, L.S., & Sprouse, J. (2015). Computational Modeling for Language Acquisition: A Tutorial With Syntactic Islands. *Journal of Speech, Language, and Hearing Research*, 740–753.

Pearl, L., & Sprouse, J. (2013). Computational models of acquisition for islands. *Experimental Syntax and Island Effects*, 109–131. https://doi.org/10.1017/cbo9781139035309.006

Pham, C., Covey, L., Gabriele, A., Aldosari, S., & Fiorentino, R. (2020). Investigating the relationship between individual differences and island sensitivity. *Glossa: a journal of general linguistics*, *5(1)*(94), 1–17. https://doi.org/10.5334/gjgl.1199

Phi, M. (2018, September 24). *Illustrated Guide to LSTM's and GRU's: A step by step explanation*. Towards Data Science. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

Rizzi, L. (1982). Violations of the wh-island constraint in Italian and the subjacency condition. In C. Dubuisson, D. Lightfoot, & Y.C. Morin (Eds.), *Issues in Italian Syntax* (pp. 49–76). Foris.

Ross, J.R. (1967). *Infinite syntax!* Ablex.

Saeed, M. (2021, September 23). *An Introduction To Recurrent Neural Networks And The Math That Powers Them*. Machine Learning Mastery. https://machinelearningmastery.com/an-introduction-to-recurrent-neural-networks-and-the-math-that-powers-them/

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Banski, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)* (pp. 28–34). Institut für Deutsche Sprache.

Schippers, A. (2012). *Variation and change in Germanic long-distance dependencies* [Doctoral Dissertation, University of Groningen]. University of Groningen/UMCG research database.

Smith, N.J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language and Linguistic Theory*, *34*, 307–344. https://doi.org/10.1007/s11049-015-9286-8

Sprouse, J., & Hornstein, N. (2013). Experimental syntax and island effects: Toward a comprehensive theory of islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 1–18). Cambridge University Press.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A TEST OF THE RELATION BETWEEN WORKING-MEMORY CAPACITY AND SYNTACTIC ISLAND EFFECTS. *Language*, *88*, 82–123. https://www.jstor.org/stable/41348884?seq=1

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., & Berwick, R.C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, *35*(3), 575–599. https://doi.org/10.1515/tlr-2018-0005

Suijkerbuijk, M. (2021). *Unpublished raw data on the acceptability of wh-islands in Dutch and the effect of filler complexity* [Dataset]. Radboud University.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, *53*(8), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1

Walczak, S., & Cerpa, N. (2003). Artificial Neural Networks. *Encyclopedia of Physical Science and Technology*, 631–645. https://doi.org/10.1016/b0-12-227410-5/00837-1

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S., & Bowman, S.R. (2019). *BLiMP: The Benchmark of Linguistic Minimal Pairs for English*. arXiv. https://doi.org/10.48550/arXiv.1912.00582

Wilcox, E., Levy, R., & Futrell, R. (2019a). *Hierarchical Representation in Neural Language Models: Suppression and Recovery of Expectations*. arXiv. https://doi.org/10.48550/arXiv.1906.04068

Wilcox, E., Levy, R., & Futrell, R. (2019b). *What Syntactic Structures block Dependencies in RNN Language Models?* arXiv. https://doi.org/10.48550/arXiv.1905.10431

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 211–221). Association for Computational Linguistics.

Wilcox, E.G., Vani, P., & Levy, R.P. (2021). *A Targeted Assessment of Incremental Processing in Neural Language Models and Humans*. arXiv. https://arxiv.org/abs/2106.03232v1

Appendix A
Item analysis

Table A1

Mean rating on 7-point scale and corresponding standard deviation for each agreement error item. Sentence layer in which agreement error occurred also indicated.

Item number	Sentence layer with error	*M*	*SD*
44	3	2.25	1.40
45	3	2.54	1.54
46	2	2.16	1.19
47	1	2.03	1.46
48	3	2.80	1.86
49	2	2.26	1.43
50	3	2.45	1.42
51	1	2.04	1.31
52	3	2.25	1.26
53	2	2.05	1.26

Table A2

Mean rating on 7-point scale and corresponding standard deviation for each word salad item.

Item number	*M*	*SD*
54	1.32	.79
55	1.32	.88
56	1.27	.80
57	1.37	.75
58	1.28	.76
59	1.36	.84
60	1.25	.72
61	1.26	.67
62	1.21	.62
63	1.33	.91
64	1.18	.71

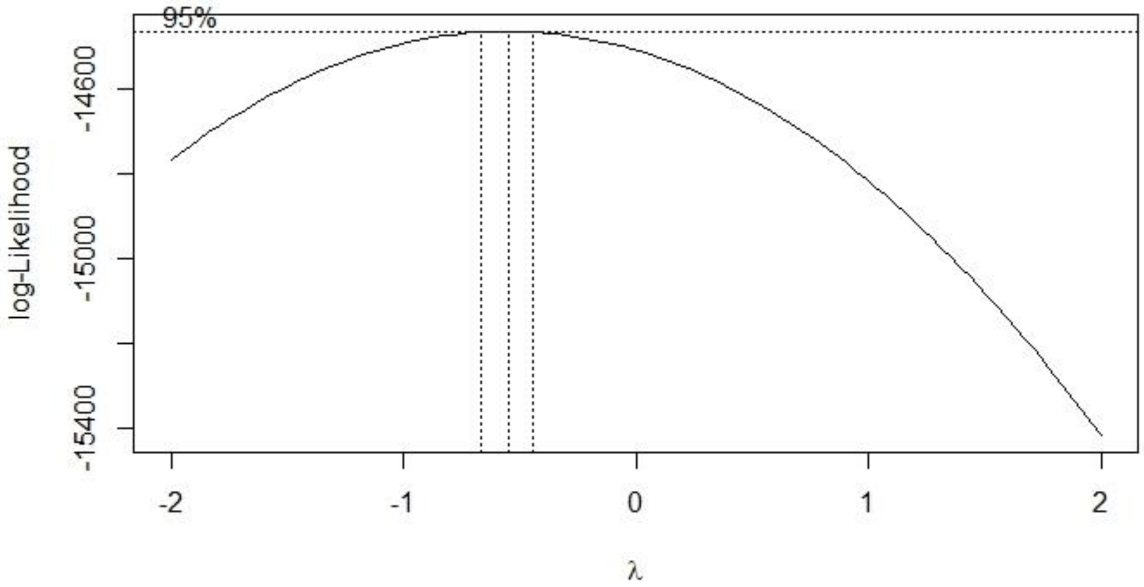
Appendix B
Model criticism (acceptability judgement task)

Analysis of experimental items

Box-cox transformation

Figure B1

Box-cox plot of data of experimental items

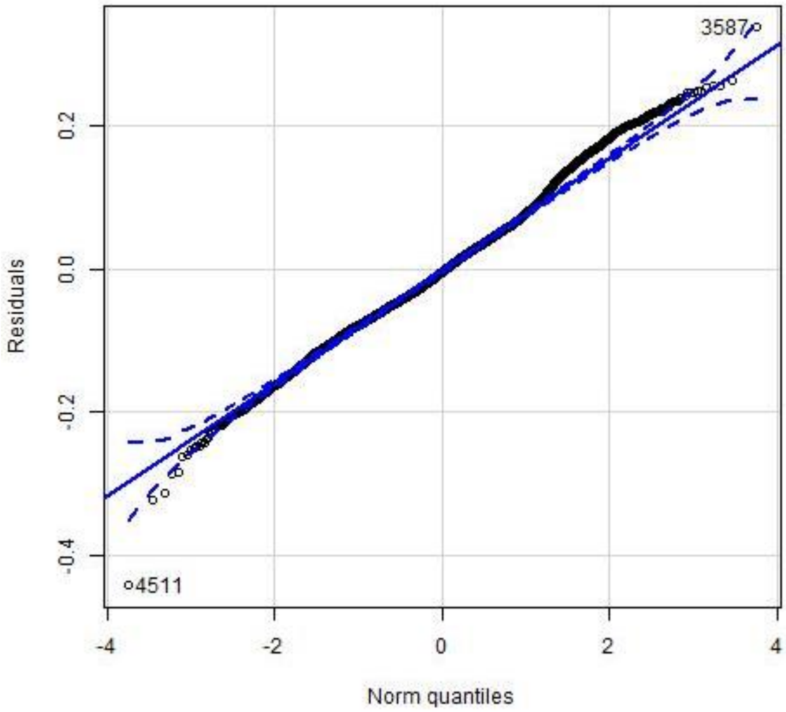


Assumption 1

The error (*ε*) must be normally distributed.

Figure B2

Q-Q plot



The box-cox transformed outcome variable roughly follows the diagonal line, which means that Assumption 1 is met.

Assumption 2

The mean of the error (*ε*) must be more or less equal to 0. This is true for the current dataset.

Assumption 3

There must be no correlation between the fitted values and the size of the error (*ε*).

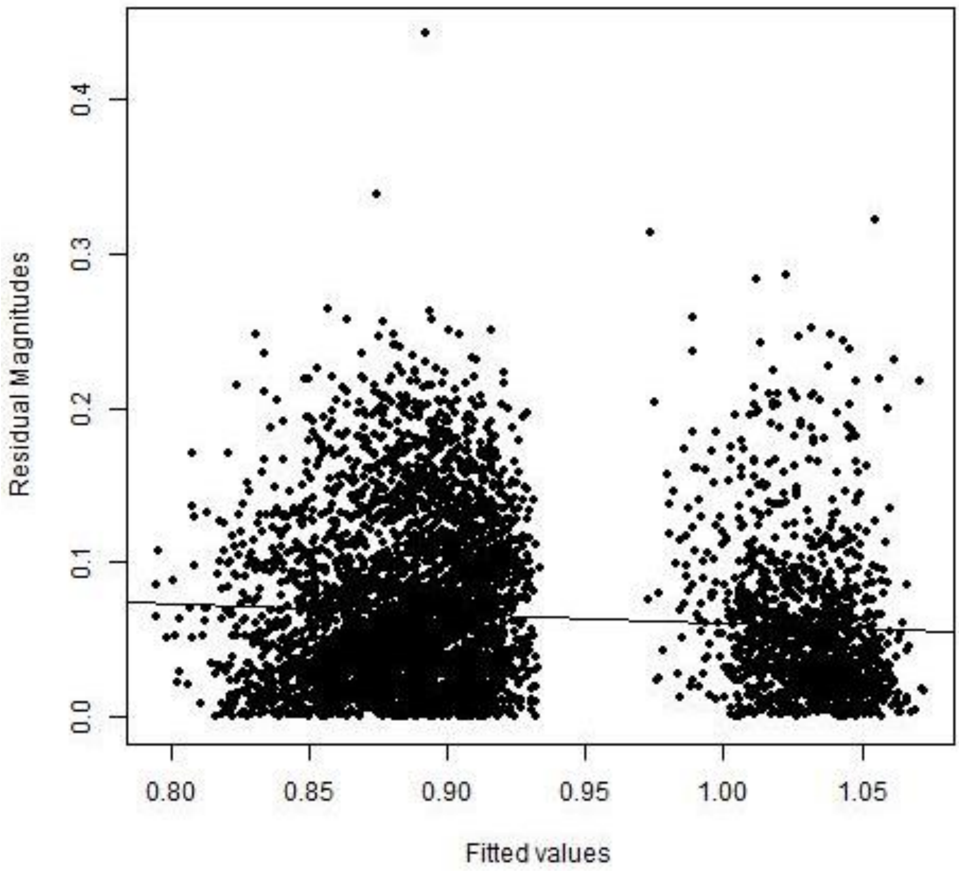
Table B1

Test of correlation with the size of the error as outcome variable and the fitted values as fixed effect.

Predictor	β [95% CI]	SE	*t*	*P*
Fitted values	−.06 [−.08, −.04]	.01	−6.06	< .001

Figure B3

Scatterplot of residual magnitudes and the fitted values.

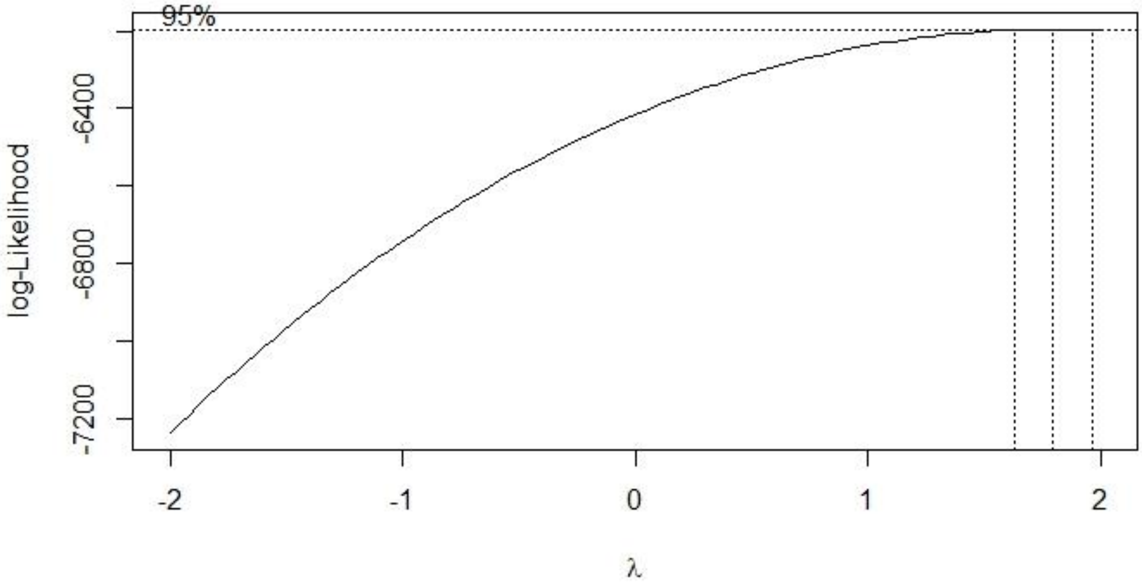


The test in Table B1 and Figure B3 show that there is correlation between the fitted values and the size of the error (ε).

Analysis of control items

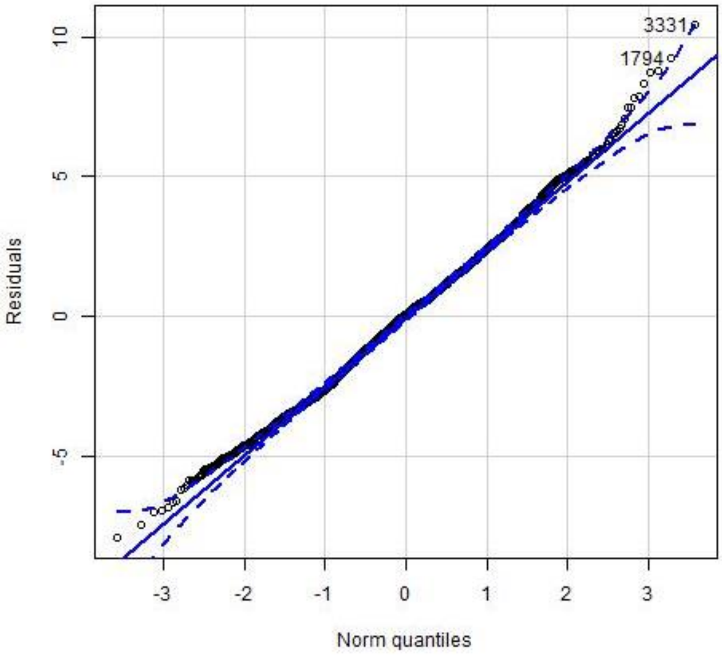
Box-cox transformation

Figure B4
Box-cox plot of data of control items



Assumption 1
The error (ε) must be normally distributed.

Figure B5
Q-Q plot



The box-cox transformed outcome variable roughly follows the diagonal line, which means that Assumption 1 is met.

Assumption 2
The mean of the error (ε) must be more or less equal to 0. This is true for the current dataset.

Assumption 3

There must be no correlation between the fitted values and the size of the error (*ε*).

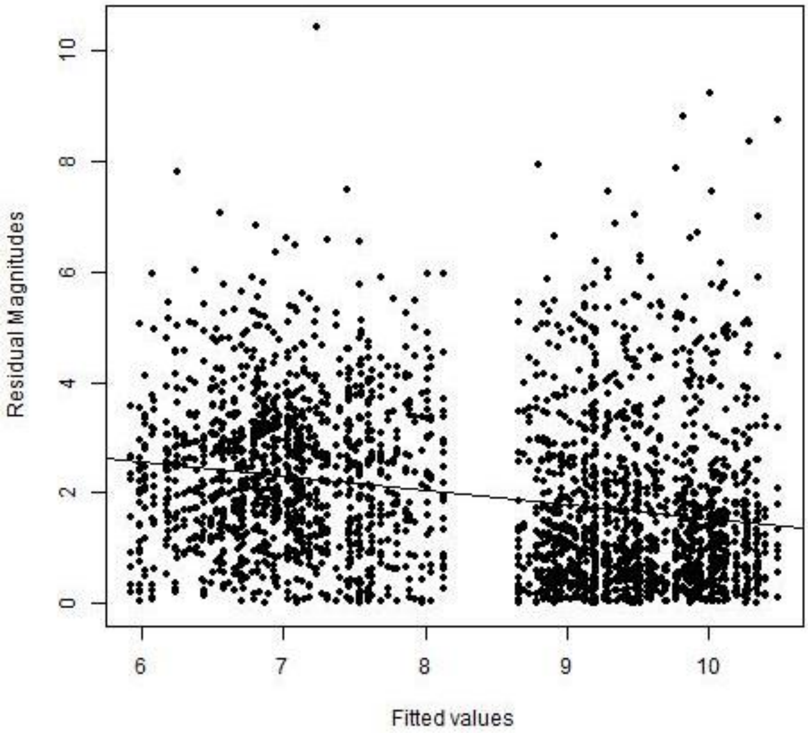
Table B2

Test of correlation with the size of the error as outcome variable and the fitted values as fixed effect.

Predictor	β [95% CI]	SE	*t*	*P*
Fitted values	−.26 [−.30, −.22]	.02	−12.88	< .001

Figure B6

Scatterplot of residual magnitudes and the fitted values.



The test in Table B2 and Figure B6 show that there is correlation between the fitted values and the size of the error (*ε*).

Appendix C
Model criticism (estimating surprisal values)

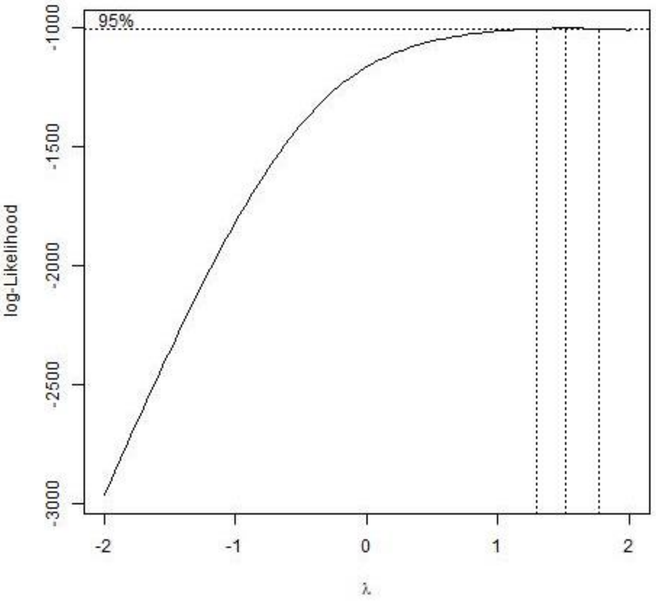
Analysis of experimental items

Box-cox transformation

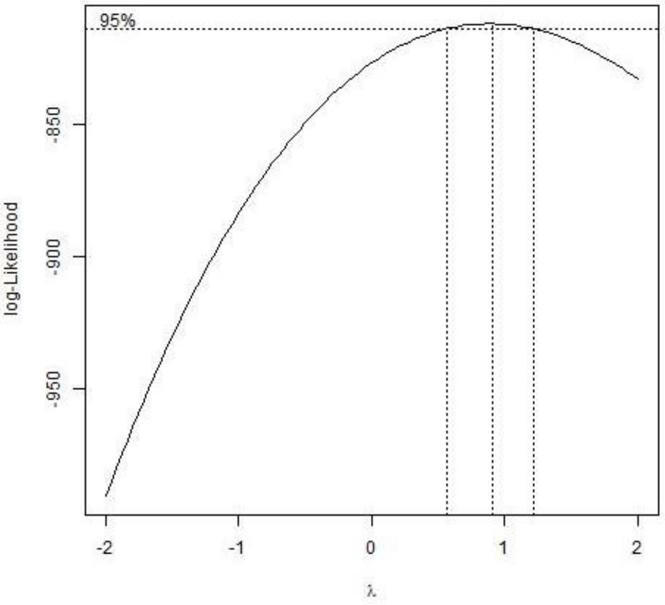
Figure C1

Box-cox plot of single-word surprisal (A) and summed surprisal (B) model.

A.



B.



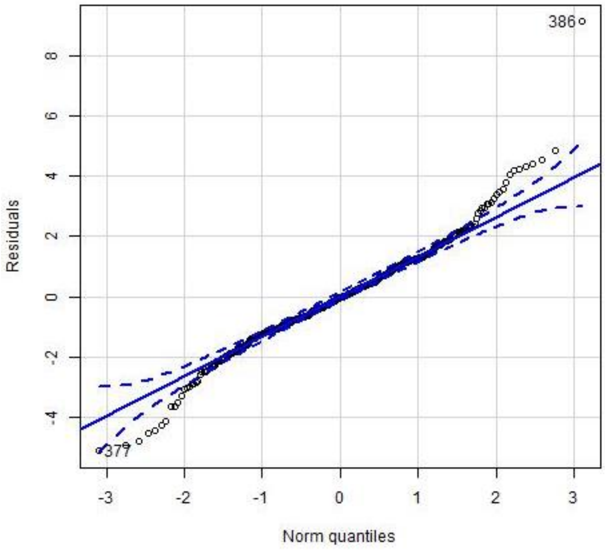
Assumption 1

The error (ε) must be normally distributed.

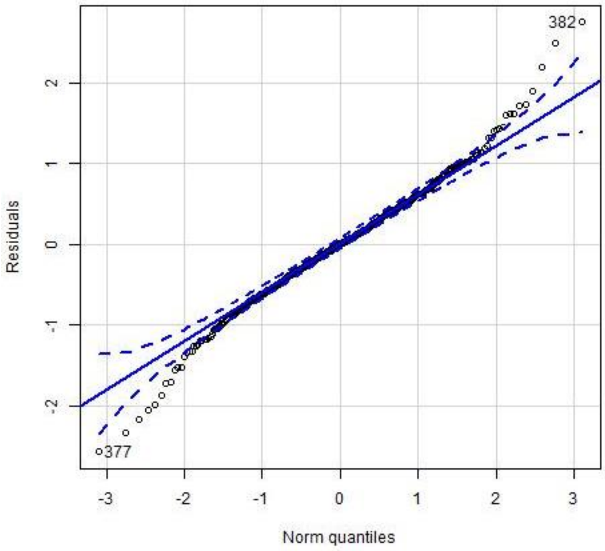
Figure C2

Q-Q plot of single-word surprisal (A) and summed surprisal (B) model.

A.



B.



The outcome variables roughly follow the diagonal line, which means that Assumption 1 is met.

Assumption 2

The mean of the error (ε) must be more or less equal to 0. This is true for both models.

Assumption 3

There must be no correlation between the fitted values and the size of the error (ε).

Table C1

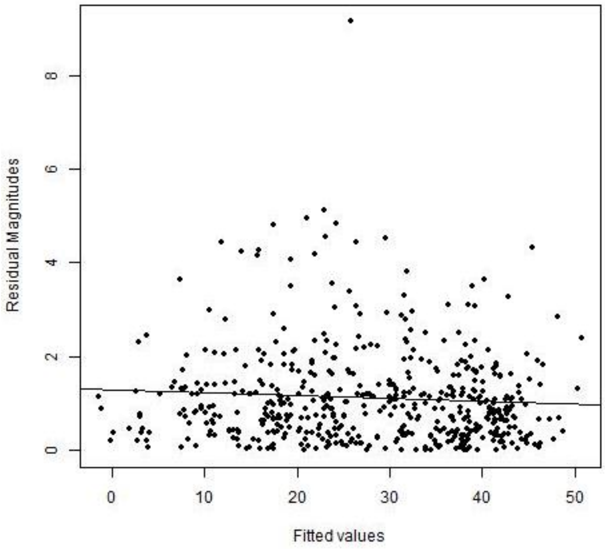
Test of correlation with the size of the error as outcome variable and the fitted values as fixed effect.

Outcome variable	Predictor	β [95% CI]	SE	*t*	*P*
Single-word surprisal	Fitted values	−.01 [−.01, .00]	.00	−1.52	.130
Summed surprisal	Fitted values	.01 [.00, .01]	.00	2.08	.038

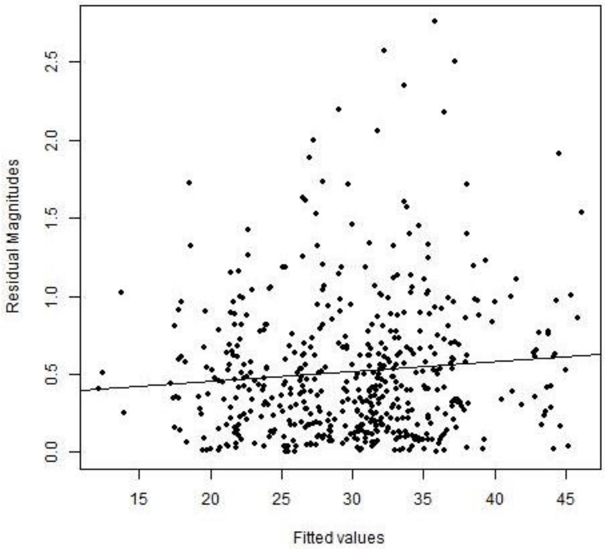
Figure C3

Scatterplot of residual magnitudes and the fitted values of the single-word surprisal (A) and summed surprisal (B) model.

A.



B.



The tests in Table C1 and Figure C3 show that there is a correlation between the fitted values and the size of the error (*ε*) for the summed surprisal model only.

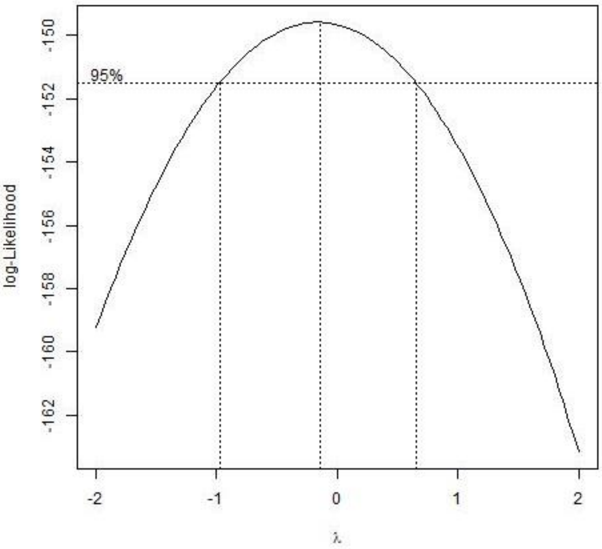
Analysis of control items

Box-cox transformation

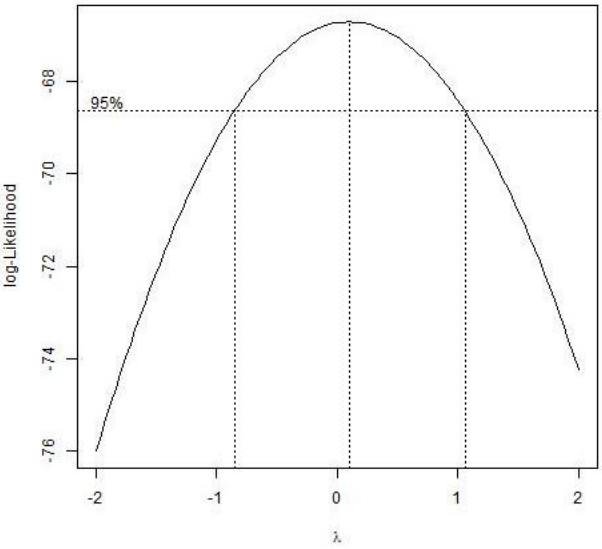
Figure C4

Box-cox plot of single-word (A) and summed (B) surprisal model.

A.



B.



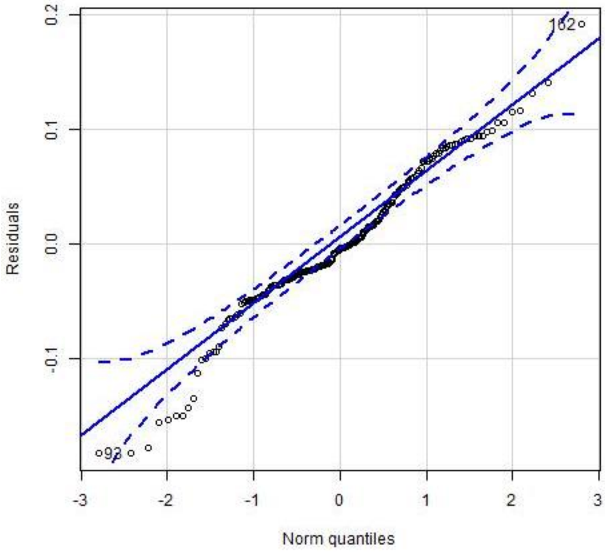
Assumption 1

The error (*ε*) must be normally distributed.

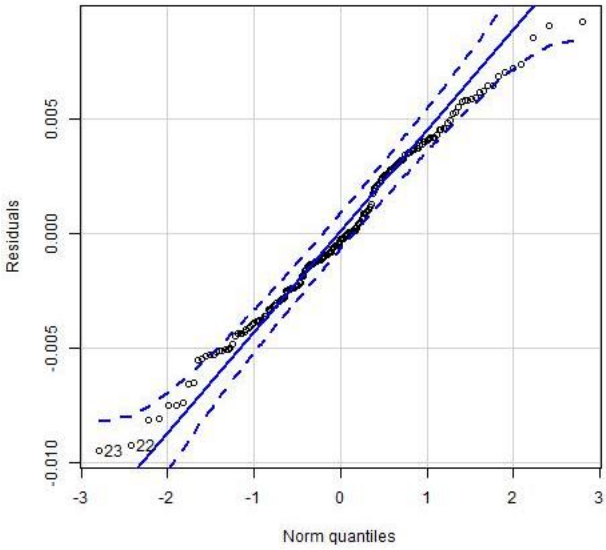
Figure C5

Q-Q plot of the single-word (A) and summed (B) surprisal model.

A.



B.



The log-transformed outcome variables roughly follow the diagonal line, which means that Assumption 1 is met.

Assumption 2

The mean of the error (ε) must be more or less equal to 0. This is true for the two models.

Assumption 3

There must be no correlation between the fitted values and the size of the error (ε).

Table C2

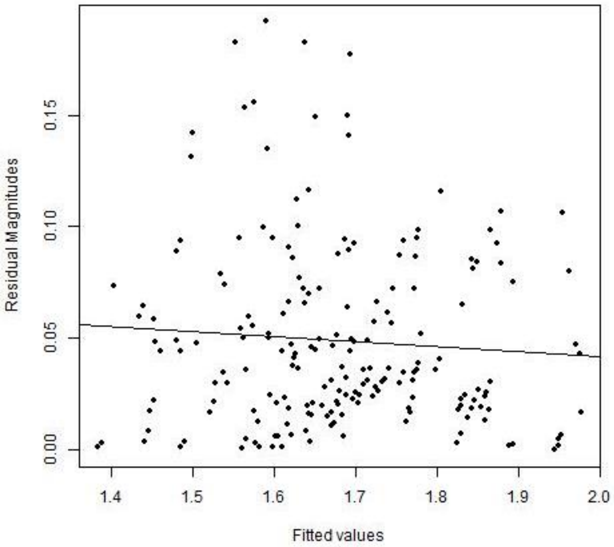
Test of correlation with the size of the error as outcome variable and the fitted values as fixed effect.

Outcome variable	Predictor	β [95% CI]	SE	*t*	*p*
Single-word surprisal	Fitted values	−.02 [−.07, .02]	.02	−1.01	.311
Summed surprisal	Fitted values	−.00 [−.01, −.00]	.00	−2.53	.012

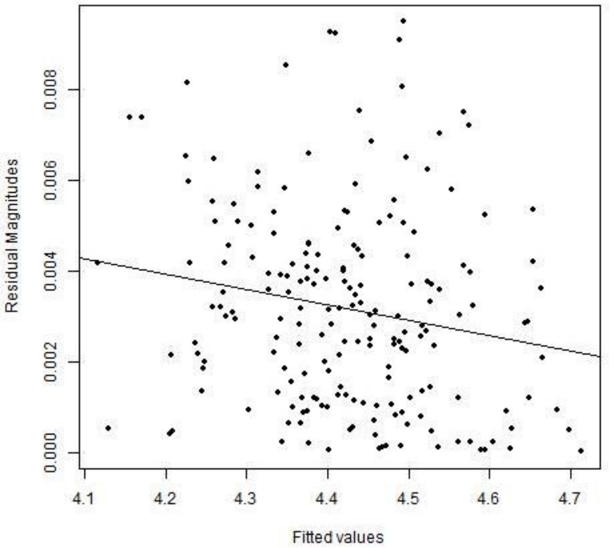
Figure C6

Scatterplot of residual magnitudes and the fitted values for the single-word (A) and summed (B) surprisal models.

A.



B.



The tests in Table C2 and Figure C7 show that there is correlation between the fitted values and the size of the error (ε) only for the summed surprisal model.

Appendix D
Frequencies of full NPs with unambiguous gender used in control items

Word	Frequency (total number of times)
meester	4,776
danser	177
docente	263
boerin	243
barman	340
actrice	732
assistente	550
redactrice	43
timmerman	568
groenteman	134
jongen	18,511
kokkin	51
ober	553
schrijfster	1,092
zangeres	941
tuinman	285
directrice	166
verkoopster	331
schoonmaakster	175
klusjesman	131
kassamedewerkster	4
postbezorgster	2
bewoner	1,803
opa	5,776
zoon	21,978
vader	41,732
zanger	1,584
acteur	1,313
verkoper	2,781
serveerster	241
juffrouw	947
danseres	128