The Learnability of the *Wh*-Island Constraint in Dutch by a Long Short-Term Memory Network

Michelle Suijkerbuijk and **Peter de Swart** and **Stefan L. Frank**
Centre for Language Studies, Radboud University
{michelle.suijkerbuijk, stefan.frank, peter.deswart}@ru.nl

Abstract

The current study investigates whether a Long Short-Term Memory (LSTM) network can learn the *wh*-island constraint in Dutch in a way comparable to human native speakers. After establishing with an acceptability judgement task that native speakers demonstrate a clear sensitivity to *wh*-island violations, the LSTM network was tested on the same sentences. Contrary to the results of the native speakers, the network was not able to recognize *wh*-islands and to block gap expectancies within them. This suggests that input and the network's inductive biases alone might not be enough to learn about syntactic island constraints, and that built-in language knowledge or abilities might be necessary.

1 Introduction

In the past decade, artificial neural networks (ANNs) have commonly been used for tasks within the research area of Natural Language Processing, such as machine translation and reading comprehension. This is a remarkable fact for many theoretical linguists, because these networks do not possess the traits considered necessary for language acquisition, such as built-in linguistic knowledge (Chomsky, 1986). Still, recent research has shown that ANNs are able to accurately learn about, for example, number agreement (i.a., Goldberg, 2019; Gulordava et al., 2018), and garden paths (i.a., Frank and Hoeks, 2019; Futrell et al., 2019; van Schijndel and Linzen, 2021). However, not all syntactic phenomena can be learned successfully yet, such as different forms of long-distance dependencies and constraints on these dependencies (Futrell et al., 2019; Wilcox et al., 2022).

One of the first computational investigations on the learnability of long-distance dependencies concerned subject-verb agreement (Gulordava et al., 2018; Linzen et al., 2016). These successful investigations showed that, when Recurrent Neural Networks (RNNs) are presented with 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'. Subject-verb agreement is a syntactic phenomenon that frequently occurs in the set of sentences the network is trained on. This makes it easy for the RNN to learn this phenomenon from only the input in combination with its inductive biases, i.e., without any built-in syntactic knowledge necessary. However, to strengthen the claim that RNNs can acquire different long-distance dependencies in this manner, it is important to also investigate dependencies not often seen in the training data set. On the one hand, if these dependencies cannot be learned by the RNN, this suggests that some built-in syntactic knowledge is necessary to learn about these long-distance dependencies. On the other hand, if the RNN can learn these dependencies, it demonstrates that the input and the network's inductive biases suffice, even if the phenomenon itself only infrequently occurs in the input. Island constraints provide an example of such an infrequent long-distance dependency and are central to the current study.

1.1 Island constraints

Filler-gap dependencies are constrained by the type of structure that can contain a gap. Previous research has shown that the filler-gap dependency in (1b) is perceived as unacceptable by most native English speakers in contrast to (1a) (Hofmeister and Sag, 2010).[1]

(1) a. *What*$_i$ did John buy $\__i$?
b. **What*$_i$ do you wonder [$_{\text{wh-phrase}}$ whether John bought $\__i$]?

[1] Gaps are represented by underscores and the *wh*-filler and gap are coindexed with i. Moreover, unacceptability is marked by an asterisk (*).

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 321-331.
Amherst, Massachusetts, June 15-17, 2023

Numerous structures (e.g., the *wh*-phrase in (1b), but also subjects, adjuncts and complex noun phrases) therefore seem to be gap-resistant (Sprouse and Hornstein, 2013; Sprouse et al., 2012). In the literature, these are referred to as *islands* (Ross, 1967), and the unacceptability caused by a filler-gap dependency in an island configuration is called an *island effect*. The current paper will focus on *wh*-islands.

There have been various investigations into the sensitivity of ANNs to the (*wh*-)island constraint, but most, if not all, focused on English. This is a problem because recent literature suggests that recurrent neural networks may have a performance advantage for English-like structural input (e.g., Dyer et al., 2019; Davis and van Schijndel, 2020), while the language learning system must be universal. Therefore, it is important to find out whether these neural networks can successfully learn about grammatical constraints such as islands in other languages as well (e.g., Kobzeva et al., 2023).

The possible performance bias for English-like structural input suggests that performance of the network will be inflated in right-branching languages such as English (i.e., with a basic word order of SVO), but undermined in left-branching and possibly mixed-branching languages (i.e., with a basic word order of SOV; Li et al., 2020).

Dutch employs mixed-branching, which means that a Dutch sentence with a matrix and an embedded clause makes use of two different branching directions; the basic and left-branching word order SOV in the embedded clause and the right-branching word order SVO in the matrix clause (due to V2; Koster, 1975). Crucially, in Dutch, the gap precedes the verb in the embedded clause, as in (2), whereas it follows the verb in English. This difference in word order due to different branching directions makes it interesting to investigate whether neural networks can learn grammatical constraints in Dutch. The current research thus focusses on Dutch as this language is typologically different from English in its word order, but shares many features as well (e.g., morphological complexity).

While there have not yet been any investigations about the performance of neural networks on island constraints in Dutch, there has been some work on the sensitivity of native speakers of Dutch to the *wh*-island constraint. Beljon et al. (2021) showed with an acceptability task that Dutch native speakers are indeed sensitive to the *wh*-island constraint. However, as this is one of only few studies to gather data on islands in Dutch, the current study will try to replicate these findings in a new acceptability judgement task. In addition, to find out whether a neural network performs comparably, a Long Short-Term Memory (LSTM) network is tested on the same sentences the speakers had to judge. The design of the test sentences was largely based on previous computational research examining island constraints in English, which we discuss below.

1.2 Island constraints and neural networks

Different computational investigations have been performed to examine whether neural networks can learn to be sensitive to island constraints. While Chowdhury and Zamparelli (2018) suggest that the networks are affected by processing factors, e.g., the syntactic complexity of islands and the position of this complex structure, Wilcox et al. (2018) argued that LSTMs can correctly learn the syntactic *wh*-, adjunct and complex noun phrase (CNP) island constraints. Wilcox et al. (2019) designed a control study to test whether a processing explanation could explain the results of Wilcox et al. (2018), and showed that LSTMs are able to learn syntactic constraints on filler-gap dependencies instead of simply being sensitive to their complexity. However, they also suggest that the networks are not completely human-like and that they are not able to learn all constraints successfully yet.

Wilcox et al. (2022) decided to combine all the knowledge gathered in these previous studies into the largest investigation to date on the network's learning ability of filler-gap dependencies and island constraints. This investigation used the same experimental design as Wilcox et al. (2018) and the control study introduced by Wilcox et al. (2019) to control for any complexity effects; we used the same design and control in the current research and will discuss them in section 2. Wilcox et al. (2022) showed that *wh*-, adjunct, CNP, left branch, and coordinate structure islands could all successfully be learned by different types of neural networks. Important to note is that these results could not be due to processing factors, as the control study used ruled out this option.

In sum, previous investigations show different results. A general agreement about whether neural networks are able to learn island constraints does thus not exist (yet), and it seems that island constraints are one of the hardest phenomena to

learn for neural networks (Warstadt et al., 2019). This makes it important to investigate why some island constraints (e.g., subject islands) are not successfully learned yet. Moreover, for the island constraints that are already successfully learned in English, it is necessary to investigate whether they can also be successfully learned in other languages. The *wh*-island constraint is, for example, successfully learned in various studies in English (e.g., Wilcox et al., 2022, 2019, 2018), 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 specifically focused on the *wh*-island constraint in Dutch.

2 Methods

To investigate the performance of the native speakers and the LSTM network on the *wh*-island constraint, we constructed experimental and control items that the speakers judged in an acceptability judgement task and that the network assigned surprisal values to.[2] Both the speakers and the network were presented with exactly the same sentences to optimize the comparison.

2.1 Experimental design

The experimental design in the current study was largely based on the interaction design introduced in Wilcox et al. (2018). This interaction design is based on two predictions assumed to be made by the grammar: (1) gaps require fillers, and (2) fillers require gaps. Consequently, the independent variables PRESENCE OF GAP and PRESENCE OF FILLER were crossed, for example in (2) for regular filler-gap dependencies.

(2) Ik weet (**wat/dat**) jij zag dat de bakker
I know (what/that) you saw that the baker
(**koekjes/_**) maakte in de bakkerij.
(cookies/GAP) made in the bakery
'I know (what/that) you saw that the baker made (cookies/_) in the bakery.'

If Dutch speakers indeed assume that fillers require gaps, filled argument positions (*koekjes* 'cookies' in (2)) should be less acceptable and more surprising when a *wh*-filler (*wat* 'what' in (2)) is present. Moreover, if Dutch speakers assume that gaps require fillers, gaps should be less acceptable and more surprising when no *wh*-filler (*dat* 'that' in (2)) is present.

Not only regular filler-gap dependencies were investigated, but also sentences with *wh*-island configurations. Therefore, the factor PRESENCE OF ISLAND was added into the interaction design as well, resulting in the four additional *wh*-island conditions illustrated in (3). The square brackets in (3) indicate the *wh*-island.

(3) Ik weet (**wat/dat**) jij je afvraagt
I know (what/that) you REF wonder
[of de bakker (**koekjes/_**) maakte in
whether the baker (cookies/GAP) made in
de bakkerij].
the bakery
'I know (what/that) you wonder whether the baker made (cookies/_) in the bakery.'

When the gaps and fillers appear in island configurations, the predictions change. First of all, a gap inside an island configuration should never be acceptable and it should be surprising for the network. Second, adding to the predictions made by Wilcox et al. (2018), the presence of a filler will increase the surprisal even more; a gap should not be expected 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. When this expectation is violated by not 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 and the surprisal value to increase. This effect should occur in sentences with and without gaps inside the island.

In total, 32 of these experimental item sets were made. The neural network saw all the conditions of each item set (and thus 256 experimental items in total), but each human participant saw only one condition per item set (and thus 32 experimental items in total).

2.2 Control items

As it is argued that humans and neural networks may simply not be able to thread information through syntactically complex constructions (i.e., islands; Keshev and Meltzer-Asscher, 2018; Wilcox et al., 2022, 2019), expectations for gendered pronouns were used to investigate this possibility (similar to the control study designed by Wilcox et al., 2019). To this end, the factors GENDER MATCH and PRESENCE OF ISLAND were

[2]The acceptability judgement task was preregistered. The preregistration can be accessed via https://doi.org/10.17605/OSF.IO/23TEQ

crossed, which resulted in four conditions: a match and mismatch condition for non-islands as in (4a) and for *wh*-islands as in (4b).

(4) a. Ik weet dat de
I know that the
(**meester/juffrouw**) denkt dat
(teacher.MASC/teacher.FEM) thinks that
de leerlingen **hem** begrijpen.
the students him understand
'I know that the (male teacher/female teacher) thinks that the students understand him.'

b. Ik weet dat de
I know that the
(**meester/juffrouw**) zich
(teacher.MASC/teacher.FEM) REF
afvraagt [of de leerlingen **hem**
wonders whether the students him
begrijpen].
understand
'I know that the (male teacher/female teacher) wonders whether the students understand him.'

It is predicted that the sentences in which the semantic gender of the noun phrase (e.g., *meester* (MASC) or *juffrouw* (FEM) 'teacher') matches the gender of the pronoun (*hem* 'him' or *haar* 'her') will be judged as more acceptable and will be less surprising than sentences in which these do not match. However, if there is any trouble in threading information through island configurations, an interaction is expected between GENDER MATCH and PRESENCE OF ISLAND; the gendered expectation effect, i.e., the difference between the sentences with matching and non-matching genders, will be reduced within island configurations. On the other hand, if the native speakers and neural network can work within complex structures, no interaction effect is expected to arise, meaning that the gendered expectation effect will arise in all configurations.

In total, 32 of these control item sets were made. The neural network saw all the conditions of each item set (and thus 128 control items in total), but each human participant saw only one condition per item set (and thus 32 control items in total).

2.3 Filler items

In addition to the experimental and control items, the human participants were also presented with 64 filler items covering the full range of acceptability; 21 acceptable (e.g., regular declarative statements), 22 moderately acceptable (e.g., anglicisms), and 21 unacceptable filler items (e.g., subject-verb agreement errors and word salads). The items and acceptability category (acceptable, moderately acceptable and unacceptable) were based on the filler items used in Beljon et al. (2021) and Kovač and Schoenmakers (2023). The unacceptable filler items were used in the current research to identify participants who appear not to perform the acceptability judgement task faithfully.

2.4 Acceptability judgement task

Participants were presented with 128 sentences (32 experimental, 32 control and 64 filler items) 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 3 filler items to familiarize them with the task. The experiment lasted 15 to 20 minutes and each participant received £3.00.

Ninety-three native speakers of Dutch, recruited from *Prolific*, entered the online experiment in *Qualtrics*. However, 29 were excluded from analyses; 6 because they did not complete the experiment and 23 because they rated more than 2 agreement errors and/or word salads with a rating of 4 or higher on the 7-point scale. The data of the remaining 64 participants ($M_{\text{age}}(SD) = 31.78(9.26)$; range: 20-55; 27 females and 34 males) were analysed.[3]

2.5 The neural network

One LSTM network was trained on a set of sentences extracted from the NLCOW2014 corpus, which comprises individual sentences of Dutch texts collected from the World Wide Web (Schäfer, 2015). Only the first slice, with approximately 37 million sentences, was used in the current research. First, a vocabulary was created by extracting the 20,000 most frequent words of the first slice and adding the set of word types used in the experimental, control and filler items of the current experi-

[3]This specific number of participants, 64, was based on a power analysis performed on unpublished data from a master's thesis. The thesis can be accessed via https://theses.ubn.ru.nl/items/a17d0411-2ed1-49b7-89cc-043540f94e00

ment, if these were not already in the most frequent word list. This resulted in a vocabulary consisting of 20,194 word types. Subsequently, only and all corpus sentences with only words from the vocabulary were selected from the first slice, and these served as training sentences.[4] The total set of training sentences comprised 8,940,314 sentences (144,196,081 tokens).

The LSTM network employed by Frank and Hoeks (2019) was used in the current study without any optimization of the architecture. It was trained on next-word prediction for 5 epochs. First, the words in the vocabulary went through a 300-unit word embedding layer. The word vectors were then passed to a 600-unit recurrent layer and a 300-unit non-recurrent layer. Last, the vectors were passed to the softmax output layer.

To check if the network was well-trained, 2 additional syntactic tests were performed. These tests explored whether the network learned correctly about (a) subject-verb agreement and (b) object-verb order in the embedded clause, a distinctive feature of Dutch (cf. section 1.1). Both are necessary syntactic skills for the network to be able to process a Dutch embedded sentence and any dependencies in it. These tests showed that the network learned both correctly. A more detailed discussion of the items used and the results can be found in Appendix A.

To evaluate the LSTM's performance, the surprisal values were collected that the network assigned to the words in the experimental and control sentences. For the experimental items, surprisal was measured at (a) the verb immediately following the gap or at the filled argument position, e.g., *maakt* 'makes' for sentences with a gap and *koekjes* 'cookies' for sentences without a gap in (2) and (3) (i.e., single-word surprisal values), and (b) summed over all words immediately following the gap or including the filled argument position, e.g., *maakt in de bakkerij* 'makes in the bakery' for sentences with a gap and *koekjes maakt in de bakkerij* 'made cookies in the bakery' for sentences without a gap in (2) and (3) (i.e., summed surprisal values). For the control items, following Wilcox et al. (2019), surprisal was measured summed over the entire sentence, and additionally at the critical pronoun *hem* 'him' or *haar* 'her'.

[4] Sentences with only one word or with more than 50 words, and sentences with a punctuation token that was not a period, comma, exclamation mark or question mark were excluded.

2.6 Data analysis

To compare the performance of Dutch native speakers and the LSTM network, surprisal values are compared to acceptability judgements following the suggestion in Pearl and Sprouse (2015); less probable words and sentences, and thus higher surprisal values, correspond to lower acceptability.

Before the statistical analysis, the raw acceptability judgement scores were converted to z-scores per participant using all items, to correct for individual differences in scale use. Additionally, all independent variables were coded using sum contrast coding, and a box-cox transformation was performed on the standardized judgement scores and the surprisal values so that the transformed data was as close to normally distributed as possible.

For both the analysis of the standardized scores and the (single-word and summed) surprisal values, two linear mixed-effects (LME) models were fitted; one for the experimental items and one for the control items. First, for the experimental items, one LME model was fitted to the standardized scores, one to the summed surprisal values and one to the single-word surprisal values with PRESENCE OF GAP, PRESENCE OF FILLER, PRESENCE OF ISLAND, and their interactions as fixed effects, using the *lmer* function from the *lmerTest* package (Kuznetsova et al., 2017) in R. Second, for the control items, one LME model was fitted to the standardized scores, one to the summed surprisal values and one to the single-word surprisal values with GENDER MATCH, PRESENCE OF ISLAND, and their interaction as fixed effects. The random effect structure for all models was based on the minimal Akaike Information Criterion (AIC). Significance values for the coefficients from all 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.

3 Results

3.1 *Wh*-island violations

The final model for the judgements included random intercepts for items and participants. The final model for the single-word surprisal included a random intercept and slope for the interaction between PRESENCE OF GAP and PRESENCE OF FILLER for items, and the final model for the summed surprisal only a random intercept for items.

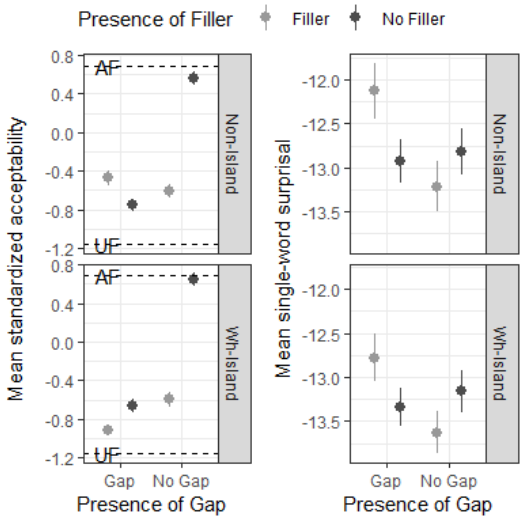

Figure 1: Mean standardized acceptability judgements (left) and mean single-word negative surprisal values (right) for every combination of PRESENCE OF GAP and PRESENCE OF FILLER for non-islands (top) and *wh*-islands (bottom). Dashed lines in the acceptability plot (left) represent the mean acceptability of the acceptable (top line; AF) and unacceptable (bottom line; UF) filler items. Error bars represent standard errors.

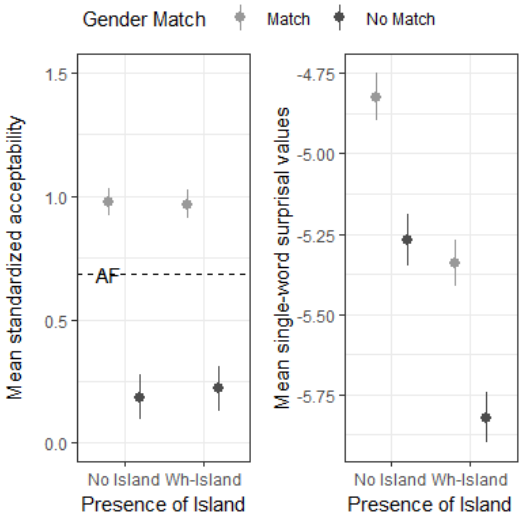

Figure 2: Mean standardized acceptability judgements (left) and mean single-word negative surprisal values (right) in non-islands and *wh*-islands with gender matches and gender mismatches. The dashed line in the acceptability plot (left) represents the mean acceptability of the acceptable filler items (AF). Error bars represent standard errors.

The results of the acceptability judgement task (left) and the LSTM network (right) are shown in Figure 1. On the y-axis of the surprisal plot, the negative surprisal values are used to facilitate the comparison with the judgement plot.

In the acceptability judgement task, a three-way interaction effect was found between PRESENCE OF GAP, PRESENCE OF FILLER, and PRESENCE OF ISLAND ($\beta = -.01$, $SE_\beta = .00$, $p < 001$). For both regular filler-gap dependencies and *wh*-islands, acceptability decreased in sentences with a filled gap when a filler was present ($M_{\text{non-island}}(SD) = -.61(.65)$, $M_{\text{island}}(SD) = -.60(.69)$) as opposed to when it was not ($M_{\text{non-island}}(SD) = .56(.63)$, $M_{\text{island}}(SD) = .65(.62)$) ($p_{\text{non-island}} < .001$, $p_{\text{island}} < .001$). However, the acceptability of regular filler-gap dependencies and *wh*-islands differed when there was a gap. In sentences with a gap, the presence of a filler increased acceptability for regular filler-gap dependencies ($M_{\text{filler}}(SD) = -.47(.70)$, $M_{\text{no filler}}(SD) = -.75(.57)$), but decreased it in a *wh*-island configuration ($M_{\text{filler}}(SD) = -.92(.45)$, $M_{\text{no filler}}(SD) = -.67(.66)$) ($p_{\text{non-island}} < .001$, $p_{\text{island}} < .001$).

For the LSTM network, no three-way interaction effect was found between PRESENCE OF GAP, PRESENCE OF FILLER, and PRESENCE OF ISLAND ($p_{\text{single-word}} = .521$, $p_{\text{summed}} = .634$), but only a two-way interaction between PRESENCE OF GAP and PRESENCE OF FILLER (single-word model: $\beta = -.16$, $SE_\beta = .02$, $p < .001$; summed model: $\beta = -.05$, $SE_\beta = .01$, $p = .002$). This means that the same patterns in surprisal were found for the regular filler-gap dependencies and *wh*-islands.[5] Specifically, surprisal increased in sentences with a filled gap when a filler was present as opposed to when it was not (non-island: $M_{\text{filler}}(SD) = 13.21(2.64)$, $M_{\text{no filler}}(SD) = 12.82(2.46)$; island: $M_{\text{filler}}(SD) = 13.63(2.27)$, $M_{\text{no filler}}(SD) = 13.16(2.21)$) ($p_{\text{non-island}} = .138$, $p_{\text{island}} = .035$), and surprisal decreased in sentences with a gap when a filler was present as opposed to when it was not (non-island: $M_{\text{filler}}(SD) = 12.13(2.96)$, $M_{\text{no filler}}(SD) = 12.92(2.34)$; island: $M_{\text{filler}}(SD) = 12.78(2.53)$, $M_{\text{no filler}}(SD) = 13.34(2.06)$) ($p_{\text{non-island}} < .001$, $p_{\text{island}} = .024$).

3.2 Gendered expectation control

The final model for the judgements included a random intercept and slope for GENDER MATCH for

[5] Only the means and standard deviations of the single-word surprisal are reported as these showed the strongest effects.

items and a random intercept for participants, and the final models for surprisal included a random intercept and slope for PRESENCE OF ISLAND for items.

The results of the participants and the LSTM network on the control items are illustrated in Figure 2. The negative surprisal values were used in the surprisal plot.

For the control items, the native speakers and the LSTM showed the same results. A main effect was found of GENDER MATCH on the standardized acceptability judgements ($\beta = 1.65$, $SE_\beta = .20$, $p < .001$) and on the summed and single-word surprisal values (single-word: $\beta = -.23$, $SE_\beta = .02$, $p < .001$; summed: $\beta = -.05$, $SE_\beta = .02$, $p = .009$), but no interaction effect between GENDER MATCH and PRESENCE OF ISLAND was found on the standardized acceptability judgements ($p = .340$) or the surprisal values ($p_{\text{single-word}} = .597$, $p_{\text{summed}} = .691$). Figure 2 shows that the sentences with a match in gender were more acceptable and less surprising than the sentences with a gender mismatch, and that this effect was the same for non-islands and islands.

4 Discussion

The current research investigated whether an LSTM network showed a similar sensitivity to *wh*-island violations in Dutch as native speakers do. After establishing whether the *wh*-island constraint exists in Dutch in an acceptability judgement task, an LSTM network was tested on the same materials and within the same experimental design to examine whether it showed similar results.

The acceptability judgement task showed that the *wh*-island constraint exists in Dutch, in line with the results by Beljon et al. (2021). Native speakers correctly showed for regular filler-gap dependencies that gaps require fillers and that fillers require gaps, and showed to be sensitive to *wh*-island violations; island configurations were only acceptable without any gaps or fillers present. These findings cannot be explained by islands being too hard to process as the control experiment showed that gender expectations could be maintained within these structures.

The network showed similar results for the regular filler-gap dependencies; it learned that gaps require fillers and that fillers require gaps. Remarkably, however, the same pattern was found within the *wh*-island configuration, contrary to the native speakers; fillers still required gaps, even when that gap then occurs within an island configuration. An LSTM network, trained on nearly 9 million Dutch sentences, does thus not seem to recognize the *wh*-island configuration in Dutch. These findings cannot be explained through processing effects, as the network could maintain gender expectations within island configurations.

While the discrepancy between human judgements and network predictions could be explained by certain design choices of the current research (e.g., the use of judgements and of complex sentences with three sentence-embedding layers), the results could also have been influenced by the architecture of the network, the training procedure, or the word order of Dutch. These factors will be discussed in turn below.

4.1 Acceptability judgements vs. surprisal

While previous research has shown that surprisal is indicative of real-time human language processing (Smith and 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. Acceptability judgements have been shown to be gradient (see Francis, 2021 for a discussion), which suggests that the knowledge underlying these judgements is probabilistic in nature instead of categorical (Lau et al., 2016). Moreover, multiple previous investigations have argued that acceptability is a concept comparable to probability, as mentioned in section 2.6 (Pearl and Sprouse, 2015; Wilcox et al., 2022). Based on this previous literature, there should be no reason to assume that the judgements and the surprisal values in the current research are not comparable.

4.2 The architecture of the network

The discrepancy between human judgements and network predictions in the current research could be explained by the specific network architecture used. While the current LSTM network does not seem successful in Dutch, other LSTM architectures have been shown to be successful in English; Wilcox et al. (2022) show that two LSTM networks can learn different island constraints successfully in English. The two LSTM networks used were the JRNN as presented in Jozefowicz et al. (2016) and the GRNN as presented in Gulordava et al. (2018). In the JRNN, the input and softmax embeddings are replaced by character convolutional neural net-

works (CNN), making it difficult to compare with the current LSTM. Moreover, the GRNN does not seem comparable either as it differs from the current LSTM in the number of hidden layers. These architectural differences could explain the results obtained for Dutch. For future research, we will thus investigate whether (a) a network more comparable to those used in Wilcox et al. (2022) for English can be successful in Dutch, and (b) the current LSTM would be successful in English.

4.3 Quantity and quality of the training data set

The difference between the human and network's results can also be due to (the size of) the data set the network is trained on. Wilcox et al. (2022) trained the GRNN on 90 million tokens and the JRNN on roughly 1 billion tokens. The current training data set comprised approximately 114 million tokens. The networks used in Wilcox et al. (2022) did not show any qualitative differences in learning success, which seems to suggest that there is no reason to believe that the size of the current data set influenced the network's learning success. While the quantity of the current training data set should thus not be of concern, the quality of the data set could have had an effect.

If the training data sets of the GRNN and the current LSTM are compared, we can identify a difference in syntactic complexity. The GRNN in Wilcox et al. (2022) was trained on English Wikipedia text, while the current training data set comprised sentences extracted from the World Wide Web. It is a well-known fact that Wikipedia text is syntactically quite complex with long and deeply embedded sentences (Yasseri et al., 2012). The current data set seems to have fewer complex sentences as, for example, more coordination conjunction is found in the longer sentences (with more than 45 words) instead of subordinating conjunction. This might mean that the number of complex sentences is smaller in the current data set than in Wikipedia text. This feature could have influenced the network's performance on the experimental items. We followed Wilcox et al. (2022) in the design of the items by using three embedding layers, which might suit Wikipedia text better in syntactic complexity. However, Wikipedia text seems less natural than the current data set, which raises the question to what extent it can be considered natural language input. Future research could use less complex experimental sentences to evaluate the network trained on the current data set, or use a data set more comparable to the one by Wilcox et al. (2022) to train the current model.

Rather than the syntactic complexity of the training data set, it could also be the case that the information in the input (training) data might just not have been good enough to learn about the *wh*-island constraint, as many syntacticians have suggested before (Chomsky, 1965; Pearl and Sprouse, 2013). This could suggest that something else is needed than just external input to learn about the *wh*-island constraint, for example some built-in language knowledge or abilities. While more research is necessary before we can say anything about the need for built-in language knowledge or abilities, our results do suggest that the domain-general learner used in the current study (i.e., the LSTM network trained on nearly 9 million Dutch sentences) is not able to recognize the *wh*-island configuration. Moreover, this domain-general learner has been shown to perform differently than the human speakers, who have been argued to have innate domain-specific knowledge about grammatical constraints (e.g., Chomsky, 1986).

4.4 The Dutch word order

The last factor that could have influenced the results of the current study is word order. The possible performance bias for English-like structural input could mean that performance can be inflated in right-branching languages such as English, but undermined in left-branching and possibly mixed-branching languages such as Dutch (Li et al., 2020). In the current research, combinations of Dutch matrix and embedded clauses were used, and thus a combination of left- and right-branching directions. Crucially, in Dutch, the gap precedes the verb in the embedded clause, which is the other way around in English. This word order difference caused by the difference in branching direction used could have affected the network's results. The current research, however, did not test this hypothesis directly. By replicating the English study by Wilcox et al. (2022), we will be able to compare the network's performance in Dutch and English directly.

In conclusion, in the current research it was shown that an LSTM network does not seem able to recognize the *wh*-island configuration in Dutch and to block gap expectancies within this configuration, unlike native speakers of Dutch. This suggests that input alone might not be enough to learn about island constraints, and that built-in language knowledge or abilities might be necessary. Moreover, it could also suggest that the mixed-branching language Dutch is, in contrast to the right-branching language English, more difficult to grasp for a neural network. Future research is needed to explore the different explanations for the current results.

The data and code can be accessed via `https://doi.org/10.17605/OSF.IO/KT3HE`.

5 Abbreviations

REF	referential pronoun
MASC	masculine
FEM	feminine

References

Maud Beljon, Dennis Joosen, Olaf Koeneman, Bram Ploum, Noëlle Sommer, Peter de Swart, and Veerle Wilms. 2021. The effect of filler complexity and context on the acceptability of wh-island violations in Dutch. *Linguistics in the Netherlands*, 38:4–20.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN Simulations of Grammaticality Judgements on Long-Distance Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 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*, pages 1979–1990. Association for Computational Linguistics.

Chris Dyer, Gábor Melis, and Phil Blunsom. 2019. A Critical Analysis of Biased Parsers in Unsupervised Parsing. *arXiv preprint*, arXiv:1909.09428.

Elaine J. Francis. 2021. *Gradient Acceptability and Linguistic Theory*. Oxford University Press.

Stefan Frank and John Hoeks. 2019. The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *Proceedings of the Cognitive Science Society*, pages 337–343. Cognitive Science Society.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT 2019*, pages 32–42. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv preprint*, arXiv:1901.05287.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of NAACL-HLT 2018*, pages 1195–1205. Association for Computational Linguistics.

Philip Hofmeister and Ivan A. Sag. 2010. Cognitive Constraints and Island Effects. *Language*, 86(2):366–415.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. *arXiv preprint*, arXiv:1602.02410.

Mayaan Keshev and Aya Meltzer-Asscher. 2018. A processing-based account of subliminal wh-island effects. *Natural Language and Linguistic Theory*, 37(2):621–657.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2023. Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian. *PsyArXiv preprint*.

Jan Koster. 1975. Dutch as an SOV language. *Linguistic Analysis*, 1:111–136.

Iva Kovač and Gert-Jan Schoenmakers. 2023. An experimental-syntactic take on long passive in Dutch: Unraveling the patterns underlying its (un)acceptability. *Manuscript submitted for publication*, University of Vienna and Radboud University.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13).

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2016. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

Russell V. Lenth. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. *R package version 1.8.3*.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2020. On the Branching Bias of Syntax Extracted from Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4473–4478. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 3:521–535.

Lisa Pearl and Jon Sprouse. 2013. Computational models of acquisition for islands. *Experimental Syntax and Island Effects*, pages 109–131.

Lisa Pearl and Jon Sprouse. 2015. Computational Modeling for Language Acquisition: A Tutorial With Syntactic Islands. *Journal of Speech, Language, and Hearing Research*, pages 740–753.

John Robert Ross. 1967. *Infinite Syntax!* Ablex.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, pages 28–34. Institut für Deutsche Sprache.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Jon Sprouse and Norbert Hornstein. 2013. Experimental syntax and island effects: Toward a comprehensive theory of islands. In *Experimental Syntax and Island Effects*, pages 1–18. Cambridge University Press.

Jon Sprouse, Matt Wagers, and Colin Phillips. 2012. A Test of the Relation Between Working-Memory Capacity and Syntactic Island Effects. *Language*, 88:82–123.

Marten van Schijndel and Tal Linzen. 2021. Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, 45(6):1–31.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *arXiv preprint*, arXiv:1912.00582.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2022. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

Ethan Gotlieb Wilcox, Roger Levy, and Richard Futrell. 2019. What Syntactic Structures Block Dependencies in RNN Language Models? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 1199–1205. Cognitive Science Society.

Ethan Gotlieb Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 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*, pages 211–221. Association for Computational Linguistics.

Taha Yasseri, András Kornai, and János Kertézs. 2012. A Practical Approach to Language Complexity: A Wikipedia Case Study. *PLoS ONE*, 7(11):1–8.

A Appendix

To investigate the performance of the LSTM network on the two additional tests, 15 item sets per phenomenon were created largely based on the item sets used in the main experiment. Each item set consisted of an acceptable and an unacceptable sentence. An example minimal pair for subject-verb agreement can be found in (5) and for object-verb order in (6).

(5) a. Hij weet dat de mevrouw dacht
he knows that the lady thought
dat de jager herten doodt tijdens de
that the hunter deer kills during the
jacht.
hunt
'He knows that the lady thought that the hunter kills deer during the hunt.'

b. *Hij weet dat de mevrouw dacht
he knows that the lady thought
dat de jagers herten doodt tijdens de
that the hunters deer kills during the
jacht.
hunt
*'He knows that the lady thought that the hunters kills deer during the hunt.'

(6) a. Ik weet dat jij denkt dat de bakker
I know that you think that the baker
koekjes maakt in de bakkerij.
cookies makes in the bakery
'I know that you think that the baker makes cookies in the bakery.'

b. *Ik weet dat jij denkt dat de bakker
I know that you think that the baker
maakt koekjes in de bakkerij.
makes cookies in the bakery
*'I know that you think that the baker makes cookies in the bakery.'

First, for subject-verb agreement, it was predicted that the network would assign higher surprisal values to the singular verb (*doodt* 'kills' in (5)) when it followed a plural subject (*jagers* 'hunters' in (5b)) than when it followed a singular subject (*jager* 'hunter' in (5a)). Second, for object-verb order, the network should assign higher surprisal values to the object-verb combination (*koekjes maakt* 'cookies makes' in (6)) when the verb incorrectly precedes the object.

For each phenomenon, an LME model was fitted to the surprisal values with ACCEPTABILITY

as fixed effect using the *lmer* function from the *lmerTest* package (Kuznetsova et al., 2017) in R. The random effect structure for both models was based on the minimal Akaike Information Criterion (AIC). Significance values for the coefficients from the models were calculated using the Satterthwaite approximation in *lmerTest* (Kuznetsova et al., 2017). The final models ultimately included a random intercept for items.

For both phenomena, a main effect of ACCEPTABILITY was found (agreement: $\beta = 1.20$, $SE_\beta = .10$, $p < .001$; order: $\beta = 1.47$, $SE_\beta = .25$, $p < .001$); the acceptable conditions (agreement: $M = 9.22$, $SD = 2.07$; order: $M = 22.57$, $SD = 4.65$) were assigned lower surprisal values than the unacceptable conditions (agreement: $M = 11.61$, $SD = 1.95$; order: $M = 25.52$, $SD = 3.60$).