LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian

Anastasia Kobzeva (anastasia.kobzeva@ntnu.no)
NTNU, Trondheim, Norway

Suhas Arehalli (suhas@jhu.edu)
Johns Hopkins University, Baltimore, MD, USA

Tal Linzen (linzen@nyu.edu)
New York University, New York, NY, USA

Dave Kush (dave.kush@utoronto.ca)
NTNU and University of Toronto, Toronto, ON, Canada

Abstract

One of the key features of natural languages is that they exhibit long-distance *filler-gap dependencies* (FGDs): In the sentence 'What do you think the pilot sent __?' the *wh*-filler *what* is interpreted as the object of the verb *sent* across multiple words. The ability to establish FGDs is thought to require hierarchical syntactic structure. However, recent research suggests that recurrent neural networks (RNNs) without specific hierarchical bias can learn complex generalizations about *wh*-questions in English from raw text data (Wilcox et al., 2018, 2019). Across two experiments, we probe the generality of this result by testing whether a long short-term memory (LSTM) RNN model can learn basic generalizations about FGDs in Norwegian. Testing Norwegian allows us to assess whether previous results were due to distributional statistics of the English input or whether models can extract similar generalizations in languages with different syntactic distributions. We also test the model's performance on two different types of FGDs: *wh*-questions and relative clauses, allowing us to determine if the model learns abstract generalizations about FGDs that extend beyond a single construction type. Results from Experiment 1 suggest that the model expects fillers to be paired with gaps and that this expectation generalizes across different syntactic positions. Results from Experiment 2 suggest that the model's expectations are largely unaffected by the increased linear distance between the filler and the gap. Our findings provide support for the conclusion that LSTM RNN's ability to learn basic generalizations about FGDs is robust across dependency type and language.

Keywords: Filler-Gap Dependencies, Neural Language Models, Norwegian, Relative Clauses, Embedded Questions

Introduction

Natural languages exhibit Filler-Gap Dependencies (FGDs) in which *filler* phrases are interpreted at later *gap* positions. Embedded questions like (1) are a type of FGD: the *wh*-filler *what* is interpreted as though it occupied the gap in the direct object position of the verb *sent* (marked with an underscore). Relative clauses (RCs) like (2) are also FGDs that include a relative pronoun (*that*) or a null operator as the filler and the head of the RC (*the present*), which is interpreted in the gap position.

(1) I know **what** the pilot sent __ to his family.

(2) I heard about **the present** that the pilot sent __ to his family.

Establishing an FGD requires abstract generalizations and representations. The well-formedness of an FGD must be described in terms of syntactic relations between the filler, the gap, and other elements in a hierarchical syntactic structure. FGDs are also potentially unbounded in length (3), which suggests that they cannot be adequately described in terms of linear predictability.

(3) I know **what** the guy from the airport said Mary saw that the pilot sent __ to his family after landing.

Despite the fact that FGDs require abstract generalizations over hierarchical representations, recent findings by Wilcox and colleagues (2018, 2019) suggest that Recurrent Neural Networks (RNNs, Elman, 1990), which are inherently sequence models without built-in biases for representing hierarchical structure, can learn FGDs and associated constraints on them. Specifically, the authors argue that Long Short-Term Memory (LSTM) RNNs (Hochreiter & Schmidhuber, 1997) that are trained with a generic language modeling objective on unannotated English text implicitly learn the distribution of acceptable FGDs in English. Their results indicate that the LSTMs could represent dependencies between fillers and gaps in multiple syntactic positions, maintain this relationship over large spans of text, and even obey complex constraints that govern where FGDs cannot be established. These results go in line with previous studies where LSTMs showed impressive results on linguistic processing tasks that require structurally-mediated dependencies, such as subject-verb agreement (Linzen et al., 2016; Gulordava et al., 2018) or auxiliary inversion (McCoy et al., 2018).

The results of Wilcox and colleagues (2018, 2019) are intriguing, but our ability to draw strong conclusions from this work about the general ability of LSTM RNNs is limited by the scope of previous experiments. First, past experiments have only investigated FGDs in English, which leaves open the question of whether the models could achieve similar success on input from languages with different distributional characteristics. Second, previous experiments only investigated one type of FGDs: *wh*-questions. It is unclear whether the success of past models should be attributed to idiosyncratic properties of (the distribution of) *wh*-questions or to a general ability of LSTMs to learn abstract generalizations about FGDs of any type.

We address this gap by exploring the ability of LSTM models to learn two types of FGDs in Norwegian: *wh*-questions and relative clauses. Norwegian is like English in that it permits FGDs across various syntactic positions, which facilitates close comparison. However, the morphosyntax of Nor-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

wegian differs from English in a number of respects, such that the distribution of cues to syntactic structure varies between the languages. For example, Norwegian is a V2 language that makes extensive use of fronting, which means that the mapping from surface word order to grammatical role is sometimes less obvious than in English. Norwegian also lacks morphological cues that might help learn syntactic dependencies, such as subject-verb agreement.

Testing RC dependencies in addition to *wh*-questions can also shed light on how abstract or general the LSTM's representations of FGDs are by testing whether success depends on specific overt lexical contingencies. *Wh*-words provide relatively unambiguous, superficial cues to the presence of a later gap. In some RCs, however, the cues are superficially ambiguous. In English, RCs can be introduced by the complementizer *that*, as seen in (2). But the complementizer *that* is also used in declarative complement clauses, where it does not license a gap (4). It also has other uses (e.g., determiner).

(4) I heard **that** the pilot sent the present to his family.

In Norwegian, the relative pronoun *som* is used in RCs as in (5). Similar to relative pronouns in English, *som* is ambiguous: it can be used as a comparative operator as in *Han er like høy som meg* 'He is as tall as me'.

(5) Jeg hørte om gaven som piloten sendte __ til
I heard about present.DEF REL pilot.DEF sent __ to
familien sin etter landing
family.DEF his after landing
'I heard about the present that the pilot sent to his family after landing.'

Such superficially ambiguous cues to the presence of a gap could potentially hinder (Gulordava et al., 2018) or improve the model's performance (Kam et al., 2008) on recognizing FGDs.

We now turn to our experiments. Experiment 1 explored whether an LSTM model can learn that fillers can be associated with gaps in different syntactic positions. Experiment 2 tested whether the model's representation of FGDs is robust to intervening material by manipulating the linear distance between the filler and the gap. To preview our results, we find that the model can represent both *wh*- and RC FGDs across different syntactic positions and can represent the FGDs across intervening material.

Methods

Language models

We trained an LSTM RNN with a language modeling objective. Such language models take a sequence of words as an input, transform it into a vector, and predict the most probable next word in that sequence using a softmax classifier over the model's vocabulary. Our model was trained on 113 million tokens of Norwegian Bokmål Wikipedia dump (Bokmål is one of the two written standards of Norwegian). Following (Gulordava et al., 2018), the model was a 2-layer LSTM with 650 hidden units in each layer and a vocabulary size of most frequent 50 000 tokens. It was trained for 40 epochs and achieved a perplexity of 30.4 on the validation set. We also trained a 5-gram model - a simple statistical model that can represent local dependencies between words within a 5-words window. This model was trained on the same corpus with Knesser-Ney smoothing and achieved a perplexity of 133.5 on the validation set. We primarily use this model as a baseline model.

Dependent variable

We investigate the model's syntactic generalizations about FGDs by looking at *surprisal*, which is the inverse log probability that the model assigns to a word given the previous context. Surprisal shows to what extent a word is unexpected given the model's probability distribution. Surprisal has been shown to correlate with incremental processing difficulty during human sentence processing (Hale, 2001; Levy, 2008).

Measuring filler-gap dependencies

Following Wilcox and colleagues (2018), we created our experimental items using a 2x2 factorial design that manipulated the presence of a filler and the presence of a gap in a sentence as in (6).

(6) She knows...

a. that the priest revealed the secret	-FILLER, -GAP
b. *that the priest revealed __	-FILLER, +GAP
c. *what the priest revealed the secret	+FILLER, -GAP
d. what the priest revealed __	+FILLER, +GAP

...in front of the guests at the party.

According to this factorial design, there should be an interaction between the presence of a filler and the presence of a gap, such that grammatical sentences with either no FGD (6-a) or a licensed FGD (6-d), should have lower surprisal values compared to ungrammatical sentences that contain an unlicensed gap (6-b), or a filler with no gap (6-c). To test for an interaction, we ran linear mixed-effects regression models with surprisal as a response variable, sum-coded conditions as predictors, and by-item random slopes (Barr et al., 2013).

When presenting our experimental results, we will collapse across two out of the four conditions by looking at pairwise differences between +FILLER and -FILLER conditions, which we call *filler effects*. There are two separate filler effects: a *filled gap effect* (Stowe, 1986, -GAP conditions) and an *unlicensed gap effect* (+GAP conditions).

The filled gap effect provides a measure of whether the presence of the filler triggers an expectation for an upcoming gap (in the earliest possible position). A filled gap effect is measured by comparing surprisal at NPs in the grammatical -FILLER, -GAP condition to the same NPs in the corresponding +FILLER, -GAP condition. If the model expects a gap after seeing a filler, it should assign a higher surprisal value to an NP in a potential gap position than it assigns to the same NP in a sentence without a filler (e.g., compare surprisal values at *the secret* in (6-c) v. (6-a)). Filled gap effects should manifest as positive differences in surprisal.

The unlicensed gap effect measures how 'surprised' the model is to find a gap in a sentence without a filler. The effect is calculated by comparing surprisal in the immediate post-gap region (i.e. *in front of* in (6)) in the +FILLER +GAP and -FILLER +GAP conditions. If the model knows that gaps must be licensed by a filler, surprisal in the post-gap region should be lower in +FILLER sentences than in -FILLER sentences. The unlicensed gap effect (the surprisal difference between conditions (6-d)-(6-b)) should be negative in such cases.

Experiment 1: Flexibility of filler-gap licensing

In Experiment 1 we test whether the models learn that fillers can license gaps in different syntactic positions. Following Wilcox and colleagues' methodology (2018), we tested both *wh*- and RC FGDs with gaps in subject, direct object, and oblique (complement of a prepositional phrase) positions in Norwegian. We present the materials and the results for each dependency type in turn.

Wh-dependencies

We created 20 test items according to a factorial design that crossed the 2x2 design in (6) with a factor that manipulated whether the gap was in subject, direct object, or oblique position as in (7), resulting in 12 conditions and 240 test sentences. Verbs were either ditransitive or transitive and accompanied by a prepositional phrase that could host a gap in oblique sentences. When the gap occurs in direct or oblique object positions in Norwegian, the structure of the sentence is the same as in English (7-b), (7-c). However, when the gap is in subject position, an expletive relative pronoun *som* is required in front of the gap in Norwegian (7-a), which could serve as an additional cue to the model for identifying the FGD.

(7) Hun vet... 'She knows...'

a. SUBJECT GAP
hvem som __ avslørte hemmeligheten foran
who REL __ revealed secret.DEF in front of
gjestene på festen
guests.DEF at party.DEF
'who revealed the secret in front of the guests at the party.'

b. DIRECT OBJECT GAP
hva presten avslørte __ foran gjestene på
what priest.DEF revealed __ in front of guests.DEF at
festen
party.DEF
'what the priest revealed __ in front of the guests at the party.'

c. OBLIQUE GAP
hvem presten avslørte hemmeligheten foran
who priest.DEF revealed secret.DEF in front of
__ på festen
__ at party.DEF
'who the priest revealed the secret in front of __ at the party.'

Figure 1 shows filler effects (differences between +FILLER and -FILLER conditions) measured in bits of surprisal (on the y-axis) by sentence region and gap position. Filled gap effects are measured at argument NPs in -GAP conditions (orange lines). Unlicensed gap effects are measured in the regions immediately following the gap for +GAP conditions (blue lines). Figure 2 compares the filled gap and unlicensed gap effects at each region of interest from the LSTM model to the baseline 5-gram model.

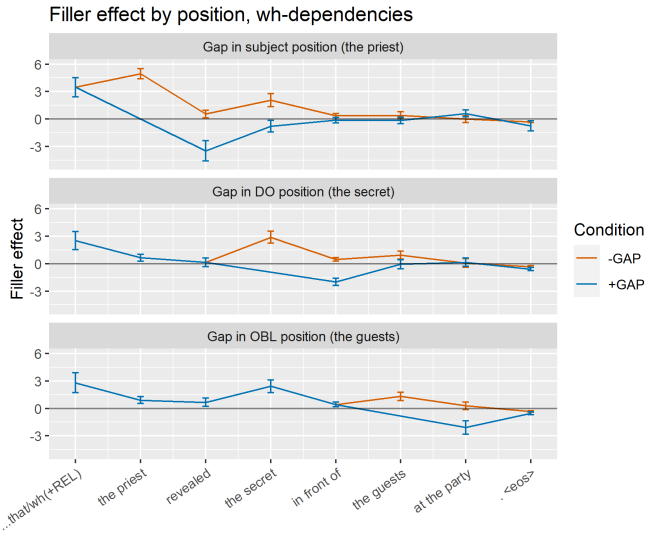

Figure 1: Filler effects for *wh*-dependencies by sentence region and gap position. Region labels are given in English for presentation purposes. Error bars are 95% confidence intervals across test items.

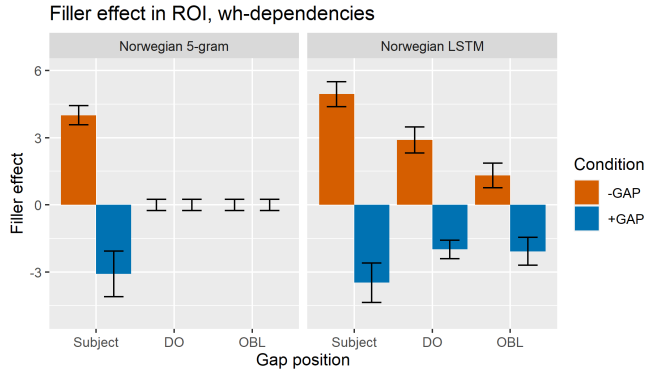

Figure 2: Filler effect for *wh*-dependencies by position.

Visual inspection of the figures suggests that the LSTM exhibits filled gap effects at all three argument positions, as evidenced by the positive surprisal differences at *the priest*, *the secret*, and *the guests*. It also appears that filled gap effects persist throughout the sentence if a gap has not been identified in an earlier position: filled gap effects are observed at the DO region after a filled subject position (top panel Figure 1) and in OBL position after a filled DO position (middle panel Figure 1). These results suggest that the model behaves like an active parser, positing a gap at every possible site after encountering a filler in the preceding context (although the effect is notably smaller in the positions following the first filled

NP position). This could be interpreted as evidence that the presence of a filler sets up an expectation for a gap in general, not in a particular syntactic position. Figure 2 also shows that the size of the filled gap effect varies by position, with subject positions inducing the largest filled gap effects, followed by direct object, and then oblique position. The baseline 5-gram model showed a filled gap effect in subject position, but nowhere else.

The model also appears to recognize unlicensed gaps in subject, DO and OBL position, as evidenced by the negative surprisal differences at *revealed*, *in front of*, and *at the party*. Once again, effects appear to be strongest in subject position, however unlicensed gap effects in DO and OBL position are comparable in size. As with the filled gap effect, the 5-gram model only exhibited an unlicensed gap effect in subject position.

Statistical analysis revealed significant interactions at all the positions tested (p <0.001 in all cases) for the LSTM. For the 5-gram model, the interaction was only significant in subject position. The fact that the 5-gram model exhibits both effects in subject position indicates that there were sentences in the training set that contained a filler and a corresponding subject gap within a 5-word window. We suspect that the apparent filled gap effects were driven by two highly frequent bigrams: *hvem som* 'who REL' and *hva som* 'what REL', where the filler is immediately adjacent to the expletive relative pronoun *som* that signals a subject gap in embedded questions. The large unlicensed gap effect can be attributed to the absence of n-grams containing the declarative complementizer *at* and the relative pronoun *som*.

Our results suggest that the LSTM model learned that *wh*-fillers can be linked to subject, object, and OBL positions in Norwegian and that gaps in these positions must be licensed by a preceding filler. Thus we replicate Wilcox and colleagues' basic findings in Norwegian. We now turn to the second part of the Experiment 1 that tested RC dependencies.

RC dependencies

The experimental items for *wh*-dependencies were modified to create sentences with RC dependencies as follows: Main-clause verbs, like *hørte* 'heard' in (8), were followed by a PP headed either by *fra* 'from' (in -FILLER) sentences or *om* 'about' (in +FILLER) sentences. PPs contained either the indefinite *noen* 'someone' or *noe* 'something'. In +FILLER sentences, the embedded clause was an RC, headed by the indefinite, followed by the relative pronoun *som*. In -FILLER sentences, the embedded clause was a complement of the main clause verb (*hørte*), followed by a PP with the indefinite *fra noen* 'from someone' and the declarative complementizer *at* 'that'. As above, the experiment manipulated the presence of a filler, the presence of a gap, and syntactic position. (8) illustrates the four SUBJECT conditions from a single item.

(8) Hun hørte... 'She heard...'

a. +FILLER, +GAP
om noen som __ avslørte hemmeligheten foran gjestene på festen
about someone REL __ revealed secret.DEF in front of guests.DEF at party.DEF
'about someone who __ revealed the secret in front of the guests at the party.'

b. +FILLER, -GAP
om noen som presten avslørte hemmeligheten foran gjestene på festen
about someone REL priest.DEF revealed secret.DEF in front of guests.DEF at party.DEF
'about someone who the priest revealed the secret in front of the guests at the party.'

c. -FILLER, +GAP
fra noen at __ avslørte hemmeligheten foran gjestene på festen
from someone that revealed secret.DEF in front of guests.DEF at party.DEF
'from someone that __ revealed the secret in front of the guests at the party.'

d. -FILLER, -GAP
fra noen at presten avslørte hemmeligheten foran gjestene på festen
from someone that priest.DEF revealed secret.DEF in front of guests.DEF at party.DEF
'from someone that the priest revealed the secret in front of the guests at the party.'

Filler effects for the LSTM model are presented in Figure 3 by sentence region and gap position. Filled gap and unlicensed gap effects for each gap position for the LSTM and the 5-gram model are in Figure 4. Overall, the qualitative pattern of effects for RC dependencies is almost identical to the pattern found with *wh*-FGDs.[1]

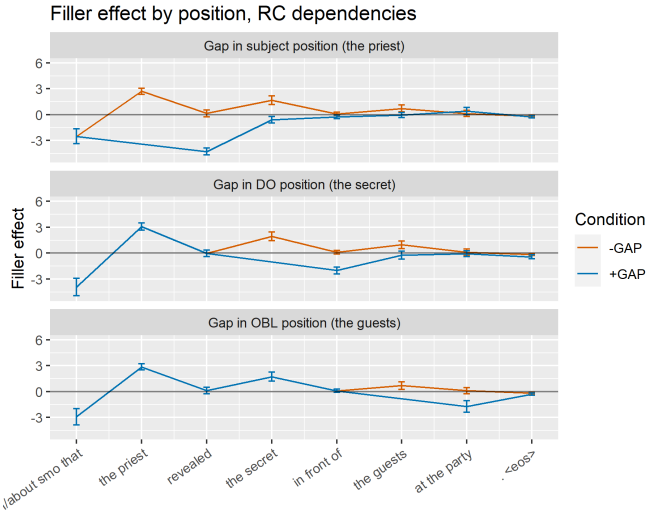

Figure 3: Filler effect for RC dependencies by sentence region and gap position.

[1]Negative difference scores in the region preceding the subject NP 'the priest' largely reflect the fact that embedding verbs like *å høre* 'to hear' are more commonly followed by the preposition *om* 'about' than by the preposition *fra* 'from'. As a result, our -FILLER sentences contained less frequent collocations in the matrix clause than +FILLER sentences, contributing to baseline surprisal differences. These differences, though, are orthogonal to our comparisons of interest.

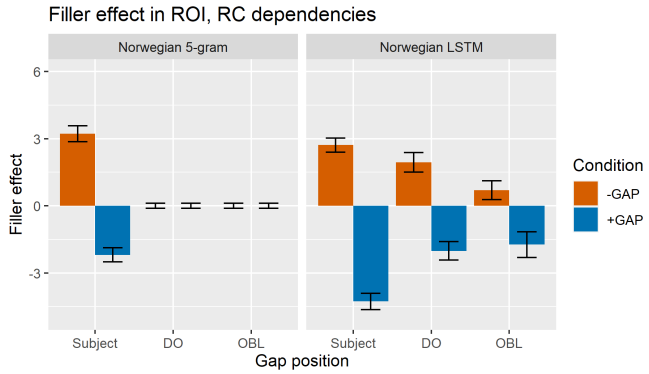

Figure 4: Filler effect for RC dependencies by position.

As with *wh*-FGDs, the LSTM model exhibits clear filled gap effects (-GAP conditions) at each potential gap position and it can distinguish between licensed and unlicensed gaps (+GAP conditions) at each position tested. Once again, the 5-gram model exhibits filled gap and unlicensed gap effects, but only in subject position. Statistical analysis confirmed a significant interaction between the presence of a filler and the presence of a gap at all three positions for the LSTM model and in subject position for the 5-gram model (p <0.001 in all cases). As with *wh*-FGDs, filled gap effects are largest in subject position and decline in size across the sentence. The unlicensed gap effect is largest in subject position, but the size does not differ between DO and OBL positions. Interestingly, the filled gap effect in subject position was larger for *wh*-FGDs (>4.5 bits) than with RCs ($\approx$ 3 bits), though the opposite was true of the unlicensed gap effect.

Experiment 2: Distance between the filler and the gap

In Experiment 2, we manipulated the linear distance between the filler and the gap to test whether the network's representation of the dependency is robust to intervening material that is irrelevant to the FGD. We manipulated distance between the filler and the gap by varying the length of a phrase modifying a subject that came between the filler and the gap, as in (Wilcox et al., 2018). As in Experiment 1, we also manipulated the presence of the filler, the presence of the gap, and the position of the gap. However, in Experiment 2 we only investigated gaps in direct object or oblique position. As in Experiment 1, we measure the size of filled gap effects and unlicensed gap effects and test whether the interaction is significant. If the model can ignore the intervening material, we expect a significant interaction between the presence of the filler and the gap at both DO and OBL positions irrespective of modifier length. If the model's ability to represent the FGD is sensitive to the intervening material, we expect a three-way interaction between the presence of the filler, the presence of the gap, and modifier length.

Wh-dependencies

We began with 20 test items crossing the presence of the filler, the presence of the gap and gap position. We crossed these items with a four-level factor controlling modifier length: No modifier as in (9-a), short modifier (2-4 words), medium modifier (5-8 words) as in (9-b), and long modifier (8-12 words), distributed across the four modifier conditions, resulting in 640 test sentences. In their original materials (Wilcox et al., 2018) used modifiers that were composed either of PPs or RCs. Our modifiers only contained PPs and conjunctions. We chose not to use RCs in our modifiers so as not to introduce any verbs that could be misinterpreted as potential gap sites between our filler and gap positions.

(9) a. NO MODIFIER
Jeg vet hva piloten sendte __ til familien sin
I know what pilot.DEF sent __ to family.DEF his
etter landing
after landing
'I know what the pilot sent __ to his family after landing.'

b. MEDIUM MODIFIER
Jeg vet hva piloten [med den blå hatten og
I know what pilot.DEF with the blue hat.DEF and
kappen] sendte __ til familien sin etter landing
coat.DEF sent __ to family.DEF his after landing
'I know what the pilot [in the blue hat and coat] sent __ to his family after landing.'

Filler effects are presented in Figure 5 by modifier and position. Not pictured are the results from the 5-gram model which showed no effects across all conditions.

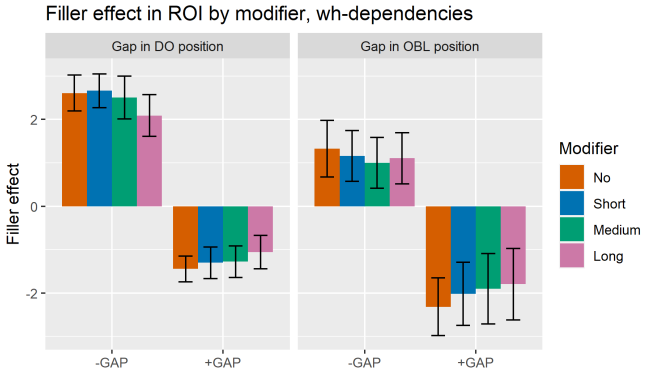

Figure 5: Filler effect for *wh*-dependencies by modifier.

As in Experiment 1, the model learned the bidirectional relationship between the presence of a filler and the presence of a gap by exhibiting filled gap effects and unlicensed gap effects in both DO and OBL position (p's <0.001). Filled gap effects were larger in DO position than in OBL position, but unlicensed gap effects were larger in OBL position. There was no significant effect of modifier length on filler-gap licensing.

RC dependencies

Materials for *wh*-FGDs were modified to create test items with RCs as in Experiment 1. Filled gap and unlicensed gap effects for RC dependencies are presented in Figure 6 by modifier and position. The 5-gram model yielded no effects.

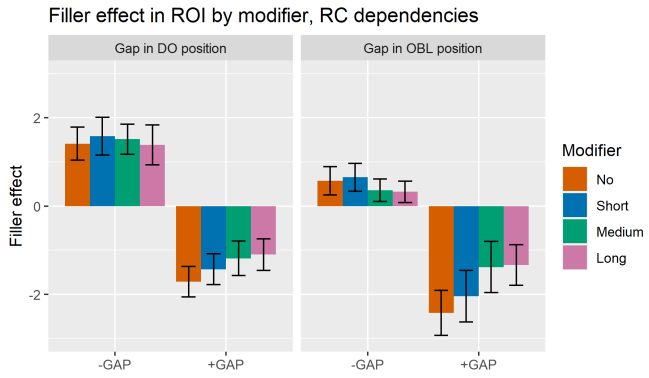

Figure 6: Filler effect for RC dependencies by modifier.

As with *wh*-dependencies, there was a significant two-way interaction between the presence of the filler and the presence of the gap for both positions (p's $<$0.001). For DO conditions, there was a significant three-way interaction between the presence of a filler, a gap and modifier length ($\beta = 0.05$, $t = 2.37$, $p = 0.018$) mostly driven by a modest diminishment in the size of the unlicensed gap effect as modifier length increased. A significant three-way interaction was also observed for OBL conditions ($\beta = 0.09$, $t = 3.44$, $p <$0.001), once again driven mostly by smaller unlicensed gap effects with longer modifiers. Despite the decrease in size, however, unlicensed gap effects are still robust across modifier length. Once again we observed that filled gap effects were rather small at the OBL position compared to the DO position.

Taken together the results of Experiment 2 suggest that the model has strong expectations for gaps in DO position with *wh*- and RC dependencies alike. Expectations for a gap in OBL position are less robust, as observed in Experiment 1, but modifier length appears to have little effect on gap expectations. The model appears to recognize unlicensed gaps in both DO and OBL position with *wh*- and RC dependencies and although the size of the effect diminishes slightly with modifier length in RC dependencies, the length of intervening material does not consistently attenuate the model's ability to detect unlicensed gaps. Overall, there seems to be an asymmetry in how the model represents the bidirectional relationship between fillers and the gaps: Unlicensed gap effects are robust and may even increase in size towards towards the end of a sentence, while filled gap effects decrease dramatically between DO and OBL position. The decrease in the size of the filled gap effect suggests that the model has weaker expectations for an RC gap in OBL position than in DO position.

Conclusions and future work

In this paper we have shown that an LSTM RNN model was able to learn two basic properties of FGDs in Norwegian: flexibility in gap position (Experiment 1) and robustness to intervening material (Experiment 2). The model appears to generalize these properties over two dependency types: *wh*- and RC dependencies. Taken together with the results of (Wilcox et al., 2018, 2019), our results provide convergent evidence that general-purpose models without pre-defined language bias can learn basic syntactic generalizations about the distribution of acceptable FGDs across different languages.

The results presented here are promising but they do not conclusively establish that the models have a robust understanding of the distribution of FGDs in Norwegian. We identify two ways in which the test materials can be modified in order to further explore the robustness of the model's generalizations. First, Experiment 2 tested the effect of *linear* distance between the filler and the gap by manipulating the length of a subject modifier phrase as in (Wilcox et al., 2018). The experiment does not establish that the model understands that FGDs are structurally unbounded, as it did not manipulate *hierarchical* distance between the filler and the gap. (Wilcox et al., 2019) showed how hierarchical distance affects the models' abilities to detect filled and unlicensed gaps in English by manipulating layers of embedding, as in (10). Future work will test the effect of hierarchical distance on Norwegian FGD licensing.

(10) I know what [the postman said [the newspaper reported [the priest revealed __ at the party]]].

Second, in both *wh*- and RC dependencies that we tested, the gaps were licensed by an *overt* lexical item. In our *wh*-FGDs the overt licensor is the *wh*-word. In our RCs the licensor was the overt relative pronoun *som*. Not all grammatical FGDs, however, require overt lexical licensing. For example, RCs without overt relative pronouns or complementizers are possible in both English and Norwegian, as shown below:

(11) a. I saw the present [$_{RC}$ the pilot sent __].
b. Jeg så gaven [$_{RC}$ piloten sendte __].
I saw present.DEF pilot.DEF sent

Testing whether the models could successfully identify licit gaps in such RCs would help determine whether the model could recognize structural cues to FGDs, or whether it was limited to lexically-signalled dependencies.

In addition to the questions mentioned above, future work will explore whether LSTMs can learn about *islands*. Islands are environments that block formation of FGDs (Ross, 1967; Chomsky et al., 1977; Huang, 1982). Wilcox et al. report that RNNs learn that *wh*-FGDs are not allowed in some island environments in English - or at least that filler-gap licensing is attenuated inside of island environments. The generality of these results should be tested in other languages. Moreover, Norwegian represents a particularly interesting case with respect to the acquisition of island constraints, because Norwegian (like other Mainland Scandinavian languages like Swedish and Danish) is argued to only exhibit sensitivity to a subset of islands that languages like English are sensitive to (Maling & Zaenen, 1982; Engdahl, 1997; Kush et al., 2021). It will be interesting to see whether RNNs can learn a different set of island constraints from different input.

Acknowledgments

Our experiments were conducted using the resources of the NTNU IDUN computing cluster (Själander, Jahre, Tufte, & Reissmann, 2019). We thank Ingrid Bondevik and Charlotte Sant for their help in creating the stimuli in Norwegian.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Chomsky, N., Culicover, P. W., Wasow, T., Akmajian, A., et al. (1977). On wh-movement. *1977*, *65*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Engdahl, E. (1997). Relative clause extractions in context. *Working Papers in Scandinavian Syntax*, *60*, 51–79.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, C. T. J. (1982). *Logical relations in Chinese and the theory of grammar* (PhD dissertation). MIT.

Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive science*, *32*(4), 771–787.

Kush, D., Sant, C., & Strætkvern, S. B. (2021). Learning island-insensitivity from the input: A corpus analysis of child- and youth-directed text in Norwegian. *Glossa: a journal of general linguistics*, *6*(1), 1–50. Retrieved 2022-01-08, from https://doi.org/10.16995/glossa.5774 doi: 10.16995/glossa.5774

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Maling, J., & Zaenen, A. (1982). A phrase structure account of Scandinavian extraction phenomena. In P. Jacobson & G. K. Pullum (Eds.), *The nature of syntactic representation* (pp. 229–282). Dordrecht: Springer Netherlands. Retrieved 2022-01-08, from https://doi.org/10.1007/978-94-009-7707-57 doi: 10.1007/978-94-009-7707-57

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

Ross, J. R. (1967). *Constraints on variables in syntax* (PhD dissertation, MIT). Retrieved 2022-01-08, from https://dspace.mit.edu/handle/1721.1/15166

Själander, M., Jahre, M., Tufte, G., & Reissmann, N. (2019). *EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure.*

Stowe, L. A. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and cognitive processes*, *1*(3), 227–245.

Wilcox, E., Levy, R., & Futrell, R. (2019). Hierarchical representation in neural language models: Suppression and recovery of expectations. *arXiv preprint arXiv:1906.04068*.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.