What can an LSTM Language Model Learn about Filler-Gap dependencies in Norwegian?

Anastasia Kobzeva[1], Suhas Arehalli[2], Tal Linzen[3], Dave Kush[1,4]

[1]Norwegian University of Science and Technology, [2]Johns Hopkins University, [3]New York University, [4]University of Toronto

Background. Recent research explores how Recurrent Neural Network (RNN) language models perform on sentence processing tasks and what kind of syntactic generalizations underlie model performance [1-3]. It has been shown that RNNs can learn *wh*-filler-gap dependencies (*wh*-FGDs) in English [4] but it is unclear yet if these results can be generalized to different languages and dependency types. In this study, we test whether RNNs can learn FGDs in Norwegian – a language that is typologically close to English – and extend the set of tested FGDs to relative clause (RC) dependencies.

Method. We trained a Long Short-Term Memory (LSTM) RNN with a language modeling objective on 113 million tokens of Norwegian Bokmål Wikipedia according to the method in [5]. We measured *surprisal* that the model assigned to words in a sentence given the previous context. We simulated how the RNN would fare as an 'incremental parser' by looking at word-by-word surprisal values in test sentences.

Following [4], we tested the model's knowledge of FGDs by probing how a filler influences the surprisal associated with (i) a later gap and (ii) the *absence* of a gap in test sentences. Experimental items followed a factorial design that manipulated the presence of a filler (+FILLER v. -FILLER) and the presence of a gap (+GAP v. -GAP). *Gap position* was also manipulated: we tested gaps in subject, direct object and oblique positions. Separate comparisons were run for *wh*- and RC-dependencies. A partial item set for *wh*-dependencies is in (1); for RC-dependencies in (2).

In one set of comparisons, we tested whether the model distinguished unlicensed gaps from licensed gaps. To do so we measured surprisal in post-gap regions in +GAP sentences (e.g., *foran* in 1c and 1d). 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 **filler effect** (the surprisal difference between conditions 1d-1c at *foran*) should be *negative* in such cases.

In a second set of comparisons, we tested whether the model exhibited filled-gap effects (FGEs) by measuring surprisal values at argument NPs in -GAP sentences (e.g., at *hemmeligheten* in 1a and 1b). If the model exhibits FGEs, it should assign higher surprisal to an NP in a potential gap position after seeing a filler than to the same NP after no filler (1b v. 1a). The **filler effect** (the surprisal difference between conditions 1b-1a at *hemmeligheten*) should be *positive* if there is an FGE.

Results and Discussion. First, we found that the model distinguishes licensed from unlicensed gaps in all positions we tested for both *wh*- and relativization dependencies, as evidenced by negative filler effects at *revealed, in front of*, and *at the party* in +GAP conditions (see blue lines in Figures 1 and 2). Second, we observed strong FGEs in subject, object, and oblique position, as evidenced by positive filler effects at corresponding NPs (see orange lines in Figures 1 and 2). The network had the strongest expectation for subject gaps, followed by DO and OBL gaps for both dependency types. Humans do not exhibit subject FGEs with *wh*-dependencies but do so with RC-dependencies [6-7]. The model differs from human performance in at least this one respect and is likely to be affected by training corpus statistics.

References. [1] Marvin, R., & Linzen, T. (2018). *EMNLP*. [2] Goldberg, Y. (2019). Assessing BERT's syntactic abilities. [3] Hu et al. (2020). *ACL*. [4] Wilcox et al. (2018). *EMNLP*. [5] Gulordava et al. (2018). *NAACL-HLT*. [6] Stowe (1986). Lang. Cogn. Process. [7]. Lee (2004). J. Psycholinguist. Res.

(1) Example stimuli set: wh-dependency (shown with gap in direct object position).

The corresponding conditions are: a. no filler, no gap; b. filler, no gap; c. no filler, gap; d. filler, gap.

a.	Hun	vet	at	presten	avslørte	hemmeligheten	foran	gjestene	på	festen.
	She	knows	that	the.priest	revealed	the.secret	in.front.of	the.guests	at	the.party.
b.	*Hun	vet	hva	presten	avslørte	hemmeligheten	foran	gjestene	på	festen.
	*She	knows	what	the.priest	revealed	the.secret	in.front.of	the.guests	at	the.party.
c.	*Hun	vet	at	presten	avslørte	___	foran	gjestene	på	festen.
	*She	knows	that	the.priest	revealed	___	in.front.of	the.guests	at	the.party.
d.	Hun	vet	hva	presten	avslørte	___	foran	gjestene	på	festen.
	She	knows	what	the.priest	revealed	___	in.front.of	the.guests	at	the.party.

(2) Example stimuli set: RC-dependency (shown with gap in direct object position).

a.	Hun	hørte	fra noen	at	presten	avslørte	hemmeligheten	foran...	gjestene...
	She	heard	from s.o.	that	the.priest	revealed	the.secret	in.front.of	the.guests
b.	*Hun	hørte	om noe	som	presten	avslørte	hemmeligheten	foran...	gjestene...
	*She	heard	about s.t.	RP	the.priest	revealed	the.secret	in.front.of	the.guests
c.	*Hun	hørte	fra noen	at	presten	avslørte	___	foran...	gjestene...
	*She	heard	from s.o.	that	the.priest	revealed	___	in.front.of	the.guests
d.	Hun	hørte	om noe	som	presten	avslørte	___	foran...	gjestene...
	She	heard	about s.t.	RP	the.priest	revealed	___	in.front.of	the.guests

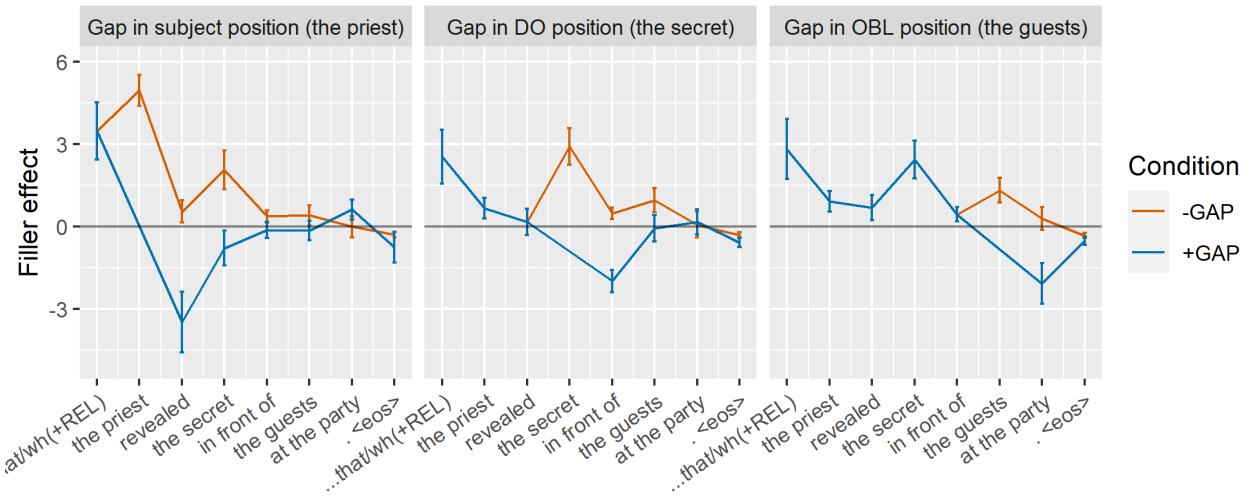

Figure 1. *Mean filler effect by region for wh-dependencies*

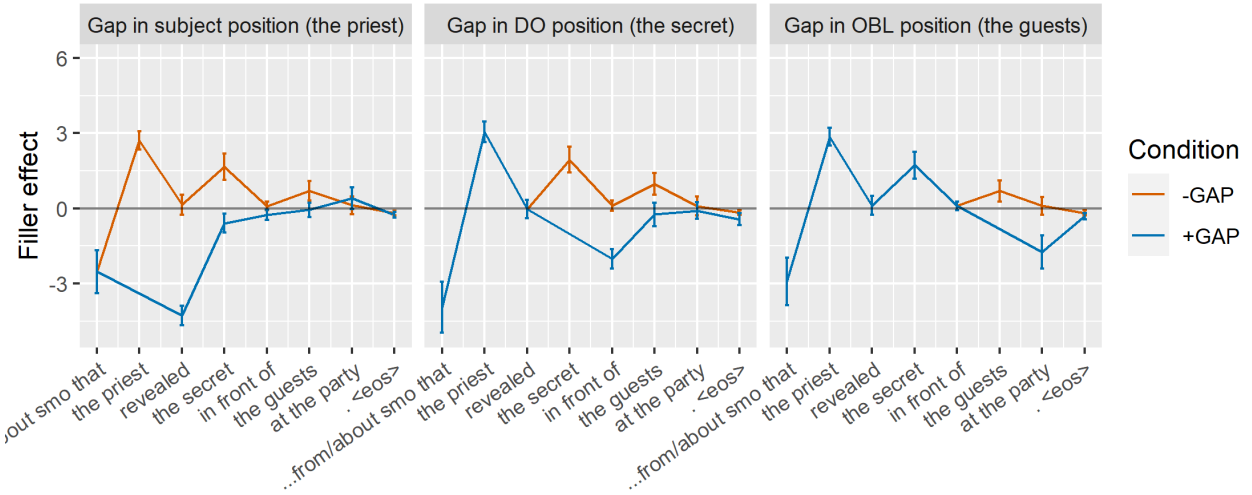

Figure 2. *Mean filler effect by region for RC-dependencies*