UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Some of Them Can be Guessed! Exploring the Effect of Linguistic Context in Predicting Quantifiers

Pezzelle, S.; Steinert Threlkeld, S.; Bernardi, R.; Szymanik, J.

DOI
10.18653/v1/P18-2019

Publication date
2018

Document Version
Final published version

Published in
ACL 2018 : The 56th Annual Meeting of the Association for Computational Linguistics

License
CC BY

Link to publication

Citation for published version (APA):
Pezzelle, S., Steinert Threlkeld, S., Bernardi, R., & Szymanik, J. (2018). *Some of* Them Can be Guessed! Exploring the Effect of Linguistic Context in Predicting Quantifiers. In I. Gurevych, & Y. Miyao (Eds.), *ACL 2018 : The 56th Annual Meeting of the Association for Computational Linguistics: proceedings of the conference : July 15-20, 2018, Melbourne, Australia* (Vol. 2, pp. 114-119). The Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2019

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

Some of Them Can be Guessed! Exploring the Effect of Linguistic Context in Predicting Quantifiers

Sandro Pezzelle[*], **Shane Steinert-Threlkeld**[†], **Raffaella Bernardi**[*‡], **Jakub Szymanik**[†]
[*]CIMeC - Center for Mind/Brain Sciences, [‡]DISI, University of Trento
[†]ILLC - Institute for Logic, Language and Computation, University of Amsterdam
[*]sandro.pezzelle@unitn.it, [†]s.n.m.steinert-threlkeld@uva.nl,
[*‡]raffaella.bernardi@unitn.it, [†]j.k.szymanik@uva.nl

Abstract

We study the role of linguistic context in predicting quantifiers ('few', 'all'). We collect crowdsourced data from human participants and test various models in a *local* (single-sentence) and a *global* context (multi-sentence) condition. Models significantly out-perform humans in the former setting and are only slightly better in the latter. While human performance improves with more linguistic context (especially on proportional quantifiers), model performance suffers. Models are very effective in exploiting lexical and morpho-syntactic patterns; humans are better at genuinely understanding the meaning of the (global) context.

1 Introduction

A typical exercise used to evaluate a language learner is the cloze deletion test (Oller, 1973). In it, a word is removed and the learner must replace it. This requires the ability to understand the context and the vocabulary in order to identify the correct word. Therefore, the larger the linguistic context, the easier the test becomes. It has been recently shown that higher-ability test takers rely more on global information, with lower-ability test takers focusing more on the local context, i.e. information contained in the words immediately surrounding the gap (McCray and Brunfaut, 2018).

In this study, we explore the role of linguistic context in predicting generalized quantifiers ('few', 'some', 'most') in a cloze-test task (see Figure 1). Both human and model performance is evaluated in a *local* (single-sentence) and a *global* context (multi-sentence) condition to study the role of context and assess the cognitive plausibility of the models. The reasons we are interested in quantifiers are myriad. First, quantifiers are of central importance in linguistic semantics and its interface with cognitive science (Barwise and Cooper, 1981; Peters and Westerståhl, 2006; Szymanik, 2016). Second, the choice of quantifier depends both on local context (e.g., positive and negative quantifiers license different patterns of anaphoric reference) and global context (the degree of positivity/negativity is modulated by discourse specificity) (Paterson et al., 2009). Third and more generally, the ability of predicting *function words* in the cloze test represents a benchmark test for human linguistic competence (Smith, 1971; Hill et al., 2016).

> ***<qnt>*** *the island's breeding birds are endemic.*
>
> *The island is one of the world's most biologically diverse areas, with many endemic species.*
> ***<qnt>*** *the island's breeding birds are endemic. Other endemic species include the red-bellied lemur, the indri, and the aye-aye.*
>
> Target quantifier: ***more than half of***

Figure 1: Given a target sentence s_t, or s_t with the preceding and following sentence, the task is to predict the target quantifier replaced by <qnt>.

We conjecture that human performance will be boosted by more context and that this effect will be stronger for *proportional* quantifiers (e.g. 'few', 'many', 'most') than for *logical* quantifiers (e.g. 'none', 'some', 'all') because the former are more dependent on discourse context (Moxey and Sanford, 1993; Solt, 2016). In contrast, we expect models to be very effective in exploiting the local context (Hill et al., 2016) but to suffer with a broader context, due to their reported inability to handle longer sequences (Paperno et al., 2016). Both hypotheses are confirmed. The best mod-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 114–119
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

els are very effective in the local context condition, where they significantly outperform humans. Moreover, model performance declines with more context, whereas human performance is boosted by the higher accuracy with proportional quantifiers like 'many' and 'most'. Finally, we show that best-performing models and humans make similar errors. In particuar, they tend to confound quantifiers that denote a similar 'magnitude' (Bass et al., 1974; Newstead and Collis, 1987).

Our contribution is twofold. First, we present a new task and results for training models to learn semantically-rich function words.[1] Second, we analyze the role of linguistic context in both humans and the models, with implications for cognitive plausibility and future modeling work.

2 Datasets

To test our hypotheses, we need linguistic contexts containing quantifiers. To ensure similarity in the syntactic environment of the quantifiers, we focus on partitive uses: where the quantifier is followed by the preposition 'of'. To avoid any effect of intensifiers like 'very' and 'so' and adverbs like 'only' and 'incredibly', we study only sentences in which the quantifier occurs at the beginning (see Figure 1). We experiment with a set of 9 quantifiers: 'a few', 'all', 'almost all', 'few', 'many', 'more than half', 'most', 'none', 'some'. This set strikes the best trade-off between number of quantifiers and their frequency in our *source* corpus, a large collection of written English including around 3B tokens.[2]

We build two datasets. One dataset – `1-Sent` – contains datapoints that only include the sentence with the quantifier (the *target* sentence, s_t). The second – `3-Sent` – contains datapoints that are 3-sentence long: the target sentence (s_t) together with both the preceding (s_p) and following one (s_f). To directly analyze the effect of the linguistic context in the task, the target sentences are exactly the same in both settings. Indeed, `1-Sent` is obtained by simply extracting all target sentences $<s_t>$ from `3-Sent` ($<s_p, s_t, s_f>$).

The `3-Sent` dataset is built as follows: (1) We split our source corpus into sentences and select those starting with a '*quantifier* of' construction. Around 391K sentences of this type are found. (2) We tokenize the sentences and replace the quantifier at the beginning of the sentence (the *target* quantifier) with the string `<qnt>`, to treat all target quantifiers as a single token. (3) We filter out sentences longer than 50 tokens (less than 6% of the total), yielding around 369K sentences. (4) We select all cases for which both the preceding and the following sentence are at most 50-tokens long. We also ensure that the target quantifier does not occur again in the target sentence. (5) We ensure that each datapoint $<s_p, s_t, s_f>$ is unique. The distribution of target quantifiers across the resulting 309K datapoints ranges from 1152 cases ('more than half') to 93801 cases ('some'). To keep the dataset balanced, we randomly select 1150 points for each quantifier, resulting in a dataset of 10350 datapoints. This was split into train (80%), validation (10%), and test (10%) sets while keeping the balancing. Then, `1-Sent` is obtained by extracting the target sentences $<s_t>$ from $<s_p, s_t, s_f>$.

3 Human Evaluation

3.1 Method

We ran two crowdsourced experiments, one per condition. In both, native English speakers were asked to pick the correct quantifier to replace `<qnt>` after having carefully read and understood the surrounding linguistic context. When more than one quantifier sounds correct, participants were instructed to choose the one they think best for the context. To make the results of the two surveys directly comparable, the same randomly-sampled 506 datapoints from the validation sets are used. To avoid biasing responses, the 9 quantifiers were presented in alphabetical order. The surveys were carried out via CrowdFlower.[3] Each participant was allowed to judge up to 25 points. To assess the judgments, 50 unambiguous cases per setting were manually selected by the native-English author and used as a benchmark. Overall, we collected judgments from 205 annotators in `1-Sent` (avg. 7.4 judgments/annotator) and from 116 in `3-Sent` (avg. 13.1). Accuracy is then computed by counting cases where at least 2 out of 3 annotators agree on the correct answer (i.e., inter-annotator agreement ≥ 0.67).

3.2 Linguistic Analysis

Overall, the task turns out to be easier in `3-Sent` (131/506 correctly-guessed cases; 0.258 accu-

[1] Data and code are at: `https://github.com/sandropezzelle/fill-in-the-quant`

[2] A concatenation of BNC, ukWaC, and a 2009-dump of Wikipedia (Baroni et al., 2014).

[3] `https://www.figure-eight.com/`

type	text	quantifier
meaning	<*qnt*> *the original station buildings survive* ***as they were used*** *as a source of materials...*	*none of*
PIs	<*qnt*> *these stories have* ***ever*** *been substantiated.*	*none of*
contrast Q	<*qnt*> *the population died out, but* ***a*** *select* ***few*** *with the right kind of genetic instability...*	*most of*
list	<*qnt*> *their major research areas are* ***social inequality, group dynamics, social change****...*	*some of*
quantity	<*qnt*> *those polled (****56%****) said that they would be willing to pay for special events...*	*more t. half of*
support Q	<*qnt*> *you have found this to be the case - click here for* ***some of*** *customer comments.*	*many of*
lexicalized	<*qnt*> ***the time****, the interest rate is set on the lender's terms...*	*most of*
syntax	<*qnt*> *these events* ***was*** *serious.*	*none of*

Table 1: Cues that might help human participants to predict the correct quantifier (`1-Sent`).

racy) compared to `1-Sent` (112/506; 0.221 acc.). Broader linguistic context is thus generally beneficial to the task. To gain a better understanding of the results, we analyze the correctly-predicted cases and look for linguistic cues that might be helpful for carrying out the task. Table 1 reports examples from `1-Sent` for each of these cues.

We identify 8 main types of cues and manually annotate the cases accordingly. (1) **Meaning**: the quantifier can only be guessed by understanding and reasoning about the context; (2) **PIs**: Polarity Items like 'ever', 'never', 'any' are licensed by specific quantifiers (Krifka, 1995); (3) **Contrast Q**: a contasting-magnitude quantifier embedded in an adversative clause; (4) **Support Q**: a supporting-magnitude quantifier embedded in a coordinate or subordinate clause; (5) **Quantity**: explicit quantitative information (numbers, percentages, fractions, etc.); (6) **Lexicalized**: lexicalized patterns like '*most of* the time'; (7) **List**: the text immediately following the quantifier is a list introduced by verbs like 'are' or 'include'; (8) **Syntax**: morpho-syntactic cues, e.g. agreement.

Figure 2 (left) depicts the distribution of annotated cues in correctly-guessed cases of `1-Sent`. Around 44% of these cases include cues besides meaning, suggesting that almost half of the cases can be possibly guessed by means of lexical factors such as PIs, quantity information, etc. As seen in Figure 2 (right), the role played by the meaning becomes much higher in `3-Sent`. Of the 74 cases that are correctly guessed in `3-Sent`, but not in `1-Sent`, more than 3 out of 4 do not display cues other than meaning. In the absence of lexical cues at the sentence level, the surrounding context thus plays a crucial role.

4 Models

We test several models, that we briefly describe below. All models except `FastText` are implemented in Keras and use `ReLu` as activation function; they are trained for 50 epochs with categorical crossentropy, initialized with frozen 300-d `word2vec` embeddings (Mikolov et al., 2013) pretrained on GoogleNews.[4] A thorough ablation study is carried out for each model to find the best configuration of parameters.[5] The best configuration is chosen based on the lowest validation loss.

BoW-conc A bag-of-words (BoW) architecture which encodes a text as the *concatenation* of the embeddings for each token. This representation is reduced by a hidden layer before softmax.

BoW-sum Same as above, but the text is encoded as the *sum* of the embeddings.

FastText Simple network for text classification that has been shown to obtain performance comparable to deep learning models (Joulin et al., 2016). `FastText` represents text as a hidden variable obtained by means of a BoW representation.

CNN Simple Convolutional Neural Network (CNN) for text classification.[6] It has two convolutional layers (`Conv1D`) each followed by `MaxPooling`. A dense layer precedes softmax.

LSTM Standard Long-Short Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997). Variable-length sequences are padded with zeros to be as long as the maximum sequence in the dataset. To avoid taking into account cells padded with zero, the 'mask zero' option is used.

bi-LSTM The Bidirectional LSTM (Schuster and Paliwal, 1997) combines information from past and future states by duplicating the first recurrent layer and then combining the two hidden states. As above, padding and mask zero are used.

[4] Available here: `http://bit.ly/1VxNC9t`

[5] We experiment with all possible combinations obtained by varying (a) optimizer: *adagrad*, *adam*, *nadam*; (b) hidden layers: 64 or 128 units; (c) dropout: 0.25, 0.5, 0.75.

[6] Adapted from: `http://bit.ly/2sFgOE1`

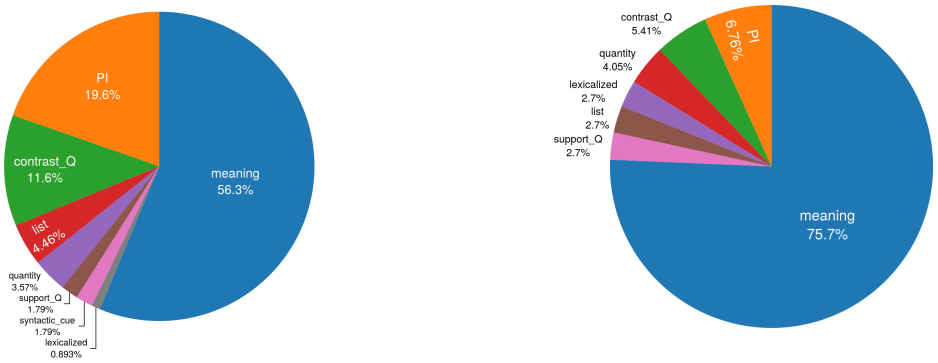

Figure 2: Left: Distribution of annotated cues across correcly-guessed cases in `1-Sent` (112 cases). Right: Distribution of cues across correctly-guessed cases in `3-Sent`, but not in `1-Sent` (74 cases).

Att-LSTM LSTM augmented with an attention mechanism (Raffel and Ellis, 2016). A feed-forward neural network computes an importance weight for each hidden state of the LSTM; the weighted sum of the hidden states according to those weights is then fed into the final classifier.

AttCon-LSTM LSTM augmented with an attention mechanism using a learned *context* vector (Yang et al., 2016). LSTM states are weighted by cosine similarity to the context vector.

5 Results

Table 2 reports the accuracy of all models and humans in both conditions. We have three main results. (1) Broader context *helps* humans to perform the task, but *hurts* model performance. This can be seen by comparing the 4-point increase of human accuracy from `1-Sent` (0.22) to `3-Sent` (0.26) with the generally worse performance of all models (e.g. `AttCon-LSTM`, from 0.34 to 0.27 in *val*). (2) All models are significantly *better* than humans in performing the task at the sentence level (`1-Sent`), whereas their performance is only slightly better than humans' in `3-Sent`. `AttCon-LSTM`, which is the best model in the former setting, achieves a significantly higher accuracy than humans' (0.34 vs 0.22). By contrast, in `3-Sent`, the performance of the best model is closer to that of humans (0.29 of `Att-LSTM` vs 0.26). It can be seen that LSTMs are overall the best-performing architectures, with `CNN` showing some potential in the handling of longer sequences (`3-Sent`). (3) As depicted in Figure 3, quantifiers that are easy/hard for humans are not necessarily easy/hard for the models. Compare 'few', 'a few', 'more than half', 'some', and 'most': while the first three are generally hard for humans but predictable by the models, the last two show the opposite pattern. Moreover, quantifiers that are guessed by humans to a larger extent in `3-Sent` compared to `1-Sent`, thus profiting from the broader linguistic context, do not experience the same boost with models. Human accuracy improves notably for 'few', 'a few', 'many', and 'most', while model performance on the same quantifiers does not.

	`1-Sent`		`3-Sent`	
	val	*test*	*val*	*test*
chance	0.111	0.111	0.111	0.111
BoW-conc	0.270	0.238	0.224	0.207
BoW-sum	0.308	0.290	0.267	0.245
fastText	0.305	0.271	0.297	0.245
CNN	0.310	0.304	**0.298**	0.257
LSTM	0.315	0.310	0.277	0.253
bi-LSTM	0.341	**0.337**	0.279	0.265
Att-LSTM	0.319	0.324	0.287	**0.291**
AttCon-LSTM	**0.343**	0.319	0.274	0.288
Humans	0.221*	——	0.258*	——

Table 2: Accuracy of models and humans. Values in **bold** are the highest in the column. *Note that due to an imperfect balancing of data, chance level for humans (computed as majority class) is 0.124.

To check whether humans and the models make similar errors, we look into the distribution of responses in `3-Sent` (*val*), which is the most comparable setting with respect to accuracy. Table 3 reports responses by humans (top) and `AttCon-LSTM` (bottom). Human errors generally involve quantifiers that display a similar magnitude as the correct one. To illustrate, 'some' is chosen in place of 'a few', and 'most' in place of either 'almost all' or 'more than half'. A similar pattern is observed in the model's predictions,

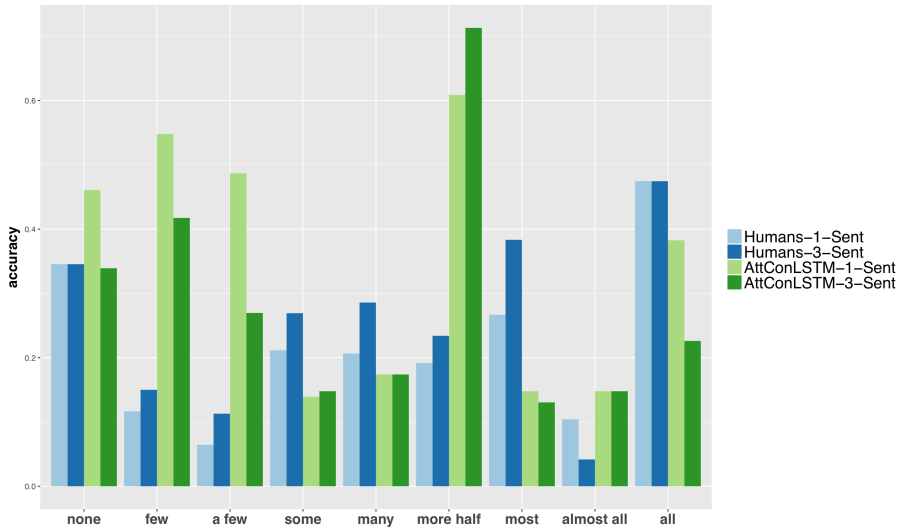

Figure 3: Human vs `AttCon-LSTM` accuracy (*val*) across quantifiers, loosely ordered by magnitude.

though we note a bias toward 'more than half'.

One last question concerns the types of linguistic cues exploited by the model (see section 3.2). We consider those cases which are correctly guessed by both humans and `AttCon-LSTM` in each setting and analyze the distribution of annotated cues. Non-semantic cues turn out to account for 41% of cases in `3-Sent` and for 50% cases in `1-Sent`. This analysis suggests that, compared to humans, the model capitalizes more on lexical, morpho-syntactic cues rather than exploiting the meaning of the context.

6 Discussion

This study explored the role of linguistic context in predicting quantifiers. For humans, the task becomes easier when a broader context is given. For the best-performing LSTMs, broader context hurts performance. This pattern mirrors evidence that predictions by these models are mainly based on local contexts (Hill et al., 2016). Corroborating our hypotheses, *proportional* quantifiers ('few', 'many', 'most') are predicted by humans with a higher accuracy with a broader context, whereas *logical* quantifiers ('all', 'none') do not experience a similar boost. Interestingly, humans are almost always able to grasp the magnitude of the missing quantifier, even when guessing the wrong one. This finding is in line with the overlapping meaning and use of these expressions (Moxey and Sanford, 1993). It also provides indirect evidence for an ordered mental scale of quantifiers (Holyoak and Glass, 1978; Routh, 1994; Moxey and Sanford, 2000). The reason why the models fail with certain quantifiers and not others is yet not clear. It may be that part of the disadvantage in the broader context condition is due to engineering issues, as suggested by an anonymous reviewer. We leave investigating these issues to future work.

none	**19**	1	2	0	2	0	0	0	12
few	5	**9**	2	6	5	0	3	0	2
a few	0	0	7	**17**	9	0	4	0	4
some	0	0	3	**14**	5	0	4	0	3
many	0	1	0	3	**18**	0	3	0	7
more than half	0	0	0	2	2	**11**	10	4	2
most	0	0	0	1	7	0	**23**	4	8
almost all	0	1	0	3	2	1	**7**	2	6
all	0	0	2	1	5	0	4	3	**28**
none	**39**	15	13	10	0	20	5	3	10
few	3	**48**	18	7	9	20	5	1	4
a few	7	13	**31**	18	5	15	12	8	6
some	5	18	16	17	16	**19**	9	5	10
many	2	18	18	15	**20**	17	10	6	9
more than half	2	7	2	3	10	**82**	2	1	6
most	8	14	14	12	12	**26**	15	5	9
almost all	5	9	15	10	8	**37**	15	6	10
all	7	12	10	15	21	13	7	4	**26**

Table 3: Responses by humans (top) and `AttCon-LSTM` (bottom) in `3-Sent` (*val*). Values in **bold** are the highest in the row.

Acknowledgments

We thank Marco Baroni, Raquel Fernández, Germán Kruszewski, and Nghia The Pham for their valuable feedback. We thank the NVIDIA Corporation for the donation of GPUs used for this research, and the iV&L Net (ICT COST Action IC1307) for funding the first author's research visit. This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 716230).

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 238–247.

Jon Barwise and Robin Cooper. 1981. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy* 4(2):159–219.

Bernard M Bass, Wayne F Cascio, and Edward J O'connor. 1974. Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology* 59(3):313.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's books with explicit memory representations. In *ICLR 2016*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Keith J Holyoak and Arnold L Glass. 1978. Recognition confusions among quantifiers. *Journal of verbal learning and verbal behavior* 17(3):249–264.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* .

Manfred Krifka. 1995. The semantics and pragmatics of polarity items. *Linguistic analysis* 25(3-4):209–257.

Gareth McCray and Tineke Brunfaut. 2018. Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing* 35(1):51–73. https://doi.org/10.1177/0265532216677105.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Linda M Moxey and Anthony J Sanford. 1993. *Communicating Quantities. A psychological perspective*. Lawrence Erlbaum Associates Publishers.

Linda M Moxey and Anthony J Sanford. 2000. Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology* 14(3):237–255.

Stephen E Newstead and Janet M Collis. 1987. Context and the interpretation of quantifiers of frequency. *Ergonomics* 30(10):1447–1462.

John W Oller. 1973. Cloze tests of second language proficiency and what they measure. *Language learning* 23(1):105–118.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of ACL 2016*.

Kevin B. Paterson, Ruth Filik, and Linda M. Moxey. 2009. Quantifiers and Discourse Processing. *Language and Linguistics Compass* .

Stanley Peters and Dag Westerståhl. 2006. *Quantifiers in Language and Logic*. Clarendon Press, Oxford.

Colin Raffel and Daniel P. W. Ellis. 2016. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. In *International Conference of Learning Representations*. http://arxiv.org/abs/1512.08756.

David A Routh. 1994. On representations of quantifiers. *Journal of Semantics* 11(3):199–214.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Frank Smith. 1971. *Understanding reading: A psycholinguistic analysis of reading and learning to read.*. Holt, Rinehart & Winston.

Stephanie Solt. 2016. On Measurement and Quantification: The Case of most and more than half. *Language* 92:65–100.

Jakub Szymanik. 2016. *Quantifiers and Cognition. Logical and Computational Perspectives*. Studies in Linguistics and Philosophy. Springer.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of NAACL-HLT 2016*. pages 1480–1489. https://doi.org/10.18653/v1/N16-1174.