Assessing Monotonicity Reasoning in Dutch through Natural Language Inference

Gijs Wijnholds
Institute for Language Sciences
Utrecht University
g.j.wijnholds@uu.nl

Abstract

In this paper we investigate monotonicity reasoning in Dutch, through a novel Natural Language Inference dataset. Monotonicity reasoning shows to be highly challenging for Transformer-based language models in English and here, we corroborate those findings using a parallel Dutch dataset, obtained by translating the Monotonicity Entailment Dataset of Yanaka et al. (2019). After fine-tuning two Dutch language models BERTje and RobBERT on the Dutch NLI dataset SICK-NL, we find that performance severely drops on the monotonicity reasoning dataset, indicating poor generalization capacity of the models. We provide a detailed analysis of the test results by means of the linguistic annotations in the dataset. We find that models struggle with downward entailing contexts, and argue that this is due to a poor understanding of negation. Additionally, we find that the choice of monotonicity context affects model performance on conjunction and disjunction. We hope that this new resource paves the way for further research in generalization of neural reasoning models in Dutch, and contributes to the development of better language technology for Natural Language Inference, specifically for Dutch.

1 Introduction

Natural Language Inference (NLI) is one of the standard benchmark tasks for current-day NLP architectures. In this task a model takes two sentences as input, and has to classify the relationship between the former (premise) sentence and the latter (hypothesis) sentence, typically between Entailment, Contradiction, and Neutral. NLI makes for a interesting task as drawing the correct inference may require subtle aspects of syntax, lexical semantics, and even pragmatics. While many NLI datasets exist like SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015) and its extensions (MNLI Williams et al. (2018), XNLI Conneau et al. (2018), e-SNLI Camburu et al. (2018)), much is still unknown about how and why neural language models (LMs) like BERT (Devlin et al., 2019) perform on the task. Evidence shows that fine-tuned LMs don't generalize well across NLI benchmarks (Talman and Chatzikyriakidis, 2019), and other investigation shows that LMs may be exploiting dataset heuristics to solve the task (Naik et al., 2018; McCoy et al., 2019). More generally, LMs do seem to encode a certain amount of syntactic structure (Rogers et al., 2020), but the relation to NLI remains unclear.

In order to shed light on the performance of large-scale LMs, specific datasets have been developed to understand what models do and don't understand. Specifically in the context of NLI, Yanaka et al. (2019) introduce the Monotonicity Entailment Dataset (MED), which targets models' capacity for understanding *monotonicity reasoning* (Icard III and Moss, 2014). Monotonicity reasoning is a staple test of human reasoning which requires lexical knowledge, as well as syntactic knowledge, making it suitable for an NLI benchmark.

In cases of monotonicity reasoning, a particular lexical item in the sentence licenses inferences by means of substituting specific syntactic constituents by either more general (upward context) or more specific (downward context).

(a) ***Every*** [*man* ↓] [*sung and danced* ↑].
(b) *Every bald man sung and danced.* ✓
(c) *Every man danced.* ✓
(d) *Every human sung and danced.* ✗

Figure 1: Example cases of monotonicity reasoning as natural language inference.

In Figure 1, the quantifier *Every* is downward entailing in its first argument, and upward entailing in its second arguments, meaning that either *man* may be replaced by a more specific instance to obtain an inference pair – as in 1(b) – while *sung and*

Findings of the Association for Computational Linguistics: EACL 2023, pages 1494–1500

danced ought to be substituted for a more general constituent to preserve inference, as in 1(c). Violating the entailment context leads to a hypothesis for which there is no entailment (but not necessarily a contradiction), as in 1(d).

While the field of research into NLI is lively, it is largely focused on English. In this article, we work with Dutch, a language that has a relatively high digital prevalence, while at the same time being underrepresented in terms of typical sentence-level NLP benchmarks.

For Dutch there is the Lassy corpus, which contains a smaller gold standard, and a larger silver standard syntactically annotated corpus of written (van Noord et al., 2013), and the SONAR corpus of written Dutch (Oostdijk et al., 2013). Given the availability of these corpora combined with a rich Wikipedia dump, two transformer-based language models have been developed for Dutch, based on the BERT architecture (BERTje, de Vries et al. (2019)) and the RoBERTa architecture (RobBERT, Delobelle et al. (2020)), both available through HuggingFace's transformers library.[1] In terms of investigations into these Dutch language models, de Vries et al. (2020) argues that BERTje encodes a typical 'NLP pipeline', which had been argued for BERT before (Tenney et al., 2019), whereas Kogkalidis and Wijnholds (2022) show through probing that long-distance dependencies are hard to recognize for both Dutch language models.

In order to extend the research done on NLI and on Dutch NLP, we add a benchmark for monotonicity reasoning in Dutch by translating the MED dataset of Yanaka et al. (2019). We perform an evaluation of large-scale language models for Dutch on this novel benchmark, that we dub MED-NL. We corroborate the findings of Yanaka et al. (2019), observing that the Dutch LMs similarly have difficulty with inferences coming from downward entailing contexts. Further inspection suggests that the main problem comes from inference pairs containing negation. In what follows, we first detail the creation of the dataset and the experimental setup for the evaluation, after which we report results and inspect the model predictions.

2 Dataset Creation & Evaluation

The dataset is obtained by translation from the English MED dataset of (Yanaka et al., 2019). First, all 5241 unique sentences are collected and lexicographically sorted to ensure consistency among sentence translations. These sentences are given to a native Dutch speaker for translation who could ensure quality and naturality of the translated examples. Using the translated sentences, we populate the original dataset with its Dutch incarnation. Since the original Entailment/Neutral labels derive from monotonicity properties, the entailment labels are preserved in Dutch. It is important to note that the labelling is binary, since MED only considers entailment and non-entailment (or neutral).

	MED	**MED-NL**
No. of tokens	81209	83809
No. of unique tokens	3614	3883
Avg. sentence length	7.54	7.79
Avg. word overlap	74.60%	73.25%

Table 1: Basic statistics of MED vs MED-NL.

Table 1 shows that in the translation, there is a 3% blowup in the number of words used in Dutch, with the corresponding increase in average sentence length. However, the number of unique tokens in the Dutch dataset increased, owing to a plurality of interpretation of English source words that may get disambiguated in the translation process.

To evaluate, we then performed a standard language model fine-tuning routine. We use two state of the art Dutch neural language models; BERTje (de Vries et al., 2019), a BERT-based model pretrained for Dutch, and RobBERT (Delobelle et al., 2020), a RoBERTa-based model for Dutch. For multilingual comparison, we furthermore train multilingual BERT (Devlin et al., 2019). Each model was trained on the SICK-NL dataset (Wijnholds and Moortgat, 2021), which is the only existing NLI benchmark for Dutch. We binarize the labels in SICK-NL by conflating all Neutral and Contradiction labels into one class, as to make the training data compatible with the binary format of MED-NL. Training proceeds for 20 epochs, and the model is saved for the epoch for which highest development accuracy is obtained.[2] We test on the SICK-NL for validation purposes, after which testing is performed on MED-NL. To reduce any potential influences of performance perturbation due

[1] There is also a distilled version of RobBERT (Delobelle et al., 2021) which we did not include in our experiments.

[2] For BERTje, highest development accuracy was achieved at epochs 3, 5, and 5, whereas RobBERT achieve peak development set performance at epochs 5, 11, and 13.

to model seed initialization, we train each model thrice and report seed-averaged accuracy.

3 Results & Analysis

Table 2 displays the average development and test accuracy on SICK-NL, and test performance on MED-NL.

	SICK-NL$_d$	SICK-NL$_t$	MED-NL
BERTje	86.89	87.40	47.56
RobBERT	86.43	85.79	46.07
mBERT	71.20	71.38	49.74

Table 2: Seed-averaged (over 3 runs) accuracy results for two Dutch BERT models and multilingual BERT, trained on SICK-NL, evaluated on both SICK-NL and the new MED-NL dataset.

Performance on the development and test set of SICK-NL are slightly higher than reported in related work (Wijnholds and Moortgat, 2021; Delobelle et al., 2021), which may be due to the fact that the classification labels have been binarized. The high drop in accuracy on MED-NL is however on par with reported results on its English counterpart (Yanaka et al., 2019), despite the models and training dataset being different. In terms of difference between the models, overall accuracy barely distinguishes BERTje and RobBERT in terms of pure performance. Interestingly, multilingual BERT has a performance decline of ca 15% compared to RobBERT, yet reached highest performance on MED-NL. The multilingual model has more trouble with the Dutch training data, although all three runs reached peak validation accuracy after one epoch of training.

Monotonicity Contexts A breakdown of accuracy results by the type of monotonicity context is given in Table 3, which shows that non-upward entailing contexts typically represent a challenge to the language models' predictions.

(*Support*)	**Total** (*5382*)	**Up** (*1818*)	**Down** (*3272*)	**Non** (*292*)
BERTje	47.56	64.76	38.72	39.50
RobBERT	46.07	61.13	39.22	28.30
mBERT	49.74	65.57	36.67	97.60

Table 3: Seed-averaged (over 3 runs) accuracy results on the MED-NL dataset, by monotonicity category.

Specifically, these results contrast the performance of the monolingual Dutch LMs with multilingual BERT, the latter doing the worst on downward entailing contexts while trumping the former models on non-monotone contexts.

Linguistic Features In order to delve deeper in the results, we make use of the annotations in the dataset that indicate specific linguistic features for premise/hypothesis pairs. Table 4 displays detailed scores for linguistic features that have a significant overall occurrence in the MED-NL dataset, where we display the number of occurrences next to the name of the feature.

	Phenomenon		**BERTje**	**RobBERT**	**mBERT**
↑	*Lexical*	743	62.72	58.73	77.39
	Conjunction	177	65.16	61.77	58.76
	Disjunction	96	24.31	29.86	53.12
	Conditionals	24	48.61	44.44	70.83
	NPI	64	33.33	36.98	64.06
	Reverse	235	52.91	51.63	50.21
	Other	698	74.79	69.91	58.74
↓	*Lexical*	477	33.47	34.45	29.98
	Conjunction	106	34.91	32.08	23.58
	Disjunction	138	49.76	49.52	40.58
	Conditionals	125	45.60	43.47	18.40
	NPI	266	36.59	39.10	32.71
	Reverse	9	29.63	33.33	33.33
	Other	2249	39.6	40.15	39.88
=	*Lexical*	182	37.73	31.32	98.35
	Disjunction	20	56.67	31.67	100.0
	NPI	8	66.67	37.50	100.0
	Other	90	39.26	23.70	95.56

Table 4: Seed-averaged (over 3 runs) accuracy results on the MED-NL dataset, by monotonicity category and phenomenon.

These results start to highlight an interesting pattern: with an overall performance on upward entailing contexts of 64.76 (BERTje), we see that cases of disjunction, conditionals, negative polarity items and reverse (e.g. double negation) are most challenging in this context. The surprising result here is that such cases are much more on par with the rest in a downward entailing context. Most strikingly, cases of disjunction become easier to deal with than conjunction in a downward entailing context, even though the situation was converse in the case of upward entailing contexts.

Model Comparison Although the results in Table 4 give some insight into the difference between models – e.g. RobBERT appears to perform higher at cases with negative polarity items, whereas BERTje performs better at cases of conjunction –, the models seem to be relatively equal in their overall accuracy. To better distinguish the models we analyse the overlap between model predictions.

	Phenomenon	∩	Shared	BERTje	RobBERT
	Lexical	75%	47.61	55.26	44.74
	Disjunction	81%	38.58	48.98	51.02
	Conjunction	82%	52.74	59.16	40.84
∀	*Conditionals*	81%	43.69	57.73	42.27
	NPI	85%	35.49	43.04	56.96
	Reverse	94%	51.6	57.21	42.79
	Other	86%	46.63	54.34	45.66
	Lexical	71%	65.05	57.47	42.53
	Disjunction	79%	20.74	39.35	60.65
	Conjunction	77%	67.56	57.69	42.31
↑	*Conditionals*	77%	45.39	63.99	36.01
	NPI	79%	31.05	39.98	60.02
	Reverse	95%	52.4	59.05	40.95
	Other	76%	79.4	60.33	39.67
	Lexical	87%	31.56	46.37	53.63
	Disjunction	86%	49.51	51.75	48.25
	Conjunction	91%	31.75	62.55	37.45
↓	*Conditionals*	82%	43.37	57.56	42.44
	NPI	87%	35.98	41.14	58.86
	Reverse	89%	27.78	33.33	66.67
	Other	90%	38.76	47.62	52.38
	Lexical	58%	23.71	56.09	43.91
	Disjunction	61%	41.55	75.14	24.86
=	*NPI*	71%	61.38	100.0	0.0
	Other	76%	26.56	75.98	24.02

Table 5: Seed-averaged overlap accuracy results on the MED-NL dataset, between BERTje and RobBERT, by monotonicity category and phenomenon.

Table 5 displays the average overlap between the two monolingual models by feature, together with their shared and individual accuracy, to shed light on where the models differ, color-coded for clarification purposes.

We first observe that the overlap between model predictions overall (the ∀ rows) is relatively high with a minimum of 75% and a maximum of 94%. Generally speaking, given that the overlap between model predictions is high, the shared accuracy shows whether models make the same correct/incorrect decisions. This is particularly pronounced in the low accuracy on disjunctions in upward entailing contexts, where models make a lot of shared mistakes, but in their diverging decisions RobBERT has a significantly higher accuracy. The converse is true for conjunction in a downward entailing context where BERT is individually stronger than RobBERT. For the sake of completeness, in Tables 8 and 9 we report overlap results between the Dutch models and multilingual BERT.

The Role of Negation One explanation for the fact that the models perform significantly worse on downward entailing contexts may be that such cases are often constructed through the use of negation words. Table 6 displays the percentages of sentence pairs containing at least one negation word, with specification for conjunction and disjunction.

	Total (*5382*)	**Up** (*1818*)	**Down** (*3272*)	**Non** (*292*)
% Negation	58.86	22.5	84.2	1.37
	↑ Conj. (*177*)	**↑ Disj.** (*96*)	**↓ Conj.** (*106*)	**↓ Disj.** (*138*)
% Negation	22.60	18.75	92.45	73.19

Table 6: Percentage of premise/hypothesis pairs in MED-NL containing negation words (*geen*, *niet*, *zonder*, *nooit*, *niemand*).

Indeed, negation is highly represented in downward monotone contexts, indicating that part of the reason why the models perform so poorly in such context is that they are not sensitive (enough) to negation. Inspection of the distribution of negation in SICK-NL (train set) and MED-NL, displayed in Table 7, shows that models may have learnt to incorrectly classify cases involving negation.

% Negation	**SICK-NL**	**MED-NL**
Entailment	1.26	69.80
Non-entailment	31.94	47.81

Table 7: Distribution of negation in cases of entailment and non-entailment in SICK-NL and MED-NL.

However, this explanation can't be replicated in the case of conjunction and disjunction, leaving a further inspection into these cases to future work.

	Phenomenon	∩	Shared	BERTje	mBERT
	Lexical	67%	60.08	28.12	71.88
	Disjunction	66%	43.03	35.47	64.53
	Conjunction	57%	49.44	59.7	40.3
∀	*Conditionals*	53%	24.23	70.3	29.7
	NPI	79%	35.37	40.16	59.84
	Reverse	91%	50.91	63.7	36.3
	Other	76%	45.67	54.2	45.8
	Lexical	60%	83.4	31.99	68.01
	Disjunction	42%	24.97	23.32	76.68
	Conjunction	41%	78.91	55.6	44.4
↑	*Conditionals*	67%	64.57	16.19	83.81
	NPI	51%	47.87	18.07	81.93
	Reverse	91%	51.72	64.94	35.06
	Other	44%	88.28	64.48	35.52
	Lexical	90%	29.57	69.73	30.27
	Disjunction	85%	44.52	78.21	21.79
	Conjunction	83%	24.94	85.0	15.0
↓	*Conditionals*	50%	13.9	77.12	22.88
	NPI	87%	32.24	66.81	33.19
	Reverse	89%	27.78	33.33	66.67
	Other	87%	38.14	50.03	49.97
	Lexical	36%	100.0	2.69	97.31
	Disjunction	57%	100.0	0.0	100.0
=	*NPI*	67%	100.0	0.0	100.0
	Other	37%	96.53	5.63	94.37

Table 8: Seed-averaged overlap accuracy results on the MED-NL dataset, between BERTje and multilingual BERT, by monotonicity category and phenomenon.

	Phenomenon	∩	Shared	RobBERT	mBERT
	Lexical	63%	58.58	27.24	72.76
	Disjunction	62%	42.34	37.75	62.25
	Conjunction	60%	46.73	56.58	43.42
∀	*Conditionals*	55%	23.38	68.16	31.84
	NPI	74%	35.7	46.86	53.14
	Reverse	90%	50.29	57.88	42.12
	Other	72%	44.75	51.0	49.0
	Lexical	59%	80.78	27.73	72.27
	Disjunction	46%	31.68	27.8	72.2
	Conjunction	44%	73.34	53.04	46.96
↑	*Conditionals*	62%	62.47	13.33	86.67
	NPI	55%	50.49	19.74	80.26
	Reverse	90%	51.01	57.88	42.12
	Other	42%	84.93	59.71	40.29
	Lexical	82%	28.22	62.06	37.94
	Disjunction	77%	43.46	70.11	29.89
	Conjunction	85%	24.01	78.27	21.73
↓	*Conditionals*	54%	14.57	76.64	23.36
	NPI	80%	32.28	65.59	34.41
	Reverse	100%	33.33	*n/a*	*n/a*
	Other	84%	38.1	50.48	49.52
	Lexical	30%	100.0	2.37	97.63
	Disjunction	32%	100.0	0.0	100.0
=	*NPI*	38%	100.0	0.0	100.0
	Other	21%	94.71	4.25	95.75

Table 9: Seed-averaged overlap accuracy results on the MED-NL dataset, between RobBERT and multilingual BERT, by monotonicity category and phenomenon.

4 Conclusion

In this paper we provided MED-NL, a novel NLI dataset for Dutch, which specifically targets monotonicity reasoning. The evaluation of two Dutch language models on this test set shows that the models specifically struggle with cases in downward entailing contexts, which had earlier been established for English as well (Yanaka et al., 2019). However, we indicate specifically that the role of negation words may play a large role in the poor model performance on such cases, giving way for future research into language models and negation.

On the other hand, the evaluation also shows that disjunction is much easier to handle by the models than conjunction, for which no explanation was found. In future investigations, we hope to provide more analysis of these language models, specifically regarding negation.

5 Limitations

This study was performed with monolingual Dutch models and with multilingual BERT, yet comparison with multilingual BERT on the original MED dataset could be insightful. Given that the distribution of cases of negation is skewed between the dataset used for training and the introduced evaluation dataset, another experiment could have been included in which models are trained to deal with cases of negation in a uniformly distributed way.

6 Acknowledgements

The author wishes to acknowledge support from the Dutch Research Council (NWO) under the scope of the project "A composition calculus for vectorbased semantic modelling with a localization for Dutch" (360-89-070). Furthermore, the author thanks Loïs Dona for help with the translation.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2021. Robbertje: A distilled dutch bert model. *Computational Linguistics in the Netherlands Journal*, 11:125–140.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas F Icard III and Lawrence S Moss. 2014. Recent progress on monotonicity. *Linguistic Issues in Language Technology*, 9:167–194.

Konstantinos Kogkalidis and Gijs Wijnholds. 2022. Discontinuous constituency and BERT: A case study of Dutch. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3776–3785, Dublin, Ireland. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 216–223, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. *The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch*, pages 219–247. Springer Berlin Heidelberg, Berlin, Heidelberg.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. *Large Scale Syntactic Annotation of Written Dutch: Lassy*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gijs Wijnholds and Michael Moortgat. 2021. SICK-NL: A dataset for Dutch natural language inference. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 1474–1479, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.