Combining Event Semantics and Degree Semantics
for Natural Language Inference

Izumi Haruta[1]
`haruta.izumi@is.ocha.ac.jp`

Koji Mineshima[2]　　　　　　　　**Daisuke Bekki**[1]
`minesima@abelard.flet.keio.ac.jp`　`bekki@is.ocha.ac.jp`

[1]Ochanomizu University, Tokyo, Japan
[2]Keio University, Tokyo, Japan

Abstract

In formal semantics, there are two well-developed semantic frameworks: event semantics, which treats verbs and adverbial modifiers using the notion of *event*, and degree semantics, which analyzes adjectives and comparatives using the notion of *degree*. However, it is not obvious whether these frameworks can be combined to handle cases in which the phenomena in question are interacting with each other. Here, we study this issue by focusing on natural language inference (NLI). We implement a logic-based NLI system that combines event semantics and degree semantics and their interaction with lexical knowledge. We evaluate the system on various NLI datasets containing linguistically challenging problems. The results show that the system achieves high accuracies on these datasets in comparison with previous logic-based systems and deep-learning-based systems. This suggests that the two semantic frameworks can be combined consistently to handle various combinations of linguistic phenomena without compromising the advantage of either framework.

1 Introduction

Since Montague (1970), formal compositional semantics has provided successful accounts of linguistic phenomena using logical expressions along with syntactic structures. In recent years, with the development of wide-coverage parsers such as Combinatory Categorial Grammar (CCG) parsers (Clark and Curran, 2007), some of the formal theories have been implemented as robust computational semantics (Bos, 2008; Mineshima et al., 2015; Abzianidze, 2016) and have been applied to natural language inference (NLI), which is the task of determining whether a text entails a hypothesis. This paper attempts to push forward this paradigm by focusing on the interaction of two semantic frameworks, namely, event semantics and degree semantics.

Generally speaking, research in linguistic formal semantics has tended to focus on creating an indepth theory for a variety of linguistic phenomena, such as quantifiers, adjectives, comparatives, and tense. However, it is often not obvious whether these independent theories can be combined and extended to cases in which the phenomena in question interact. Thus, event semantics has been developed largely to account for the semantics of verb phrases and adverbial modifiers (Davidson, 1967; Parsons, 1990). In contrast, degree semantics provides an analysis of gradable expressions such as adjectives and comparatives, using the notion of degree (Cresswell, 1976; Stechow, 1984; Kennedy, 1997). Although each theory has been elaborated for its own sake, it is not clear how to combine these two and handle expressions that require the application of both theories, such as the comparative form of adverbs, though it is a necessary step for analyzing real texts.

The computational modeling of compositional semantic theories mentioned above enables us to precisely compute the predictions of each theory. In addition, their application to NLI tasks provides a systematic way of evaluating a formal semantic theory. Building on the previous logic-based approaches to NLI, we present a logic-based NLI system that combines event semantics and degree semantics, and

Example	Logical form
John shouted *loudly*.	$\exists e(\text{shout}(e) \wedge (\text{subj}(e) = \text{john}) \wedge \text{loud}(e, \theta_{\text{loud}}))$
Ann studied English *very hard*.	$\exists e(\text{study}(e) \wedge (\text{subj}(e) = \text{ann}) \wedge (\text{obj}(e) = \text{english}) \wedge \exists \delta(\text{hard}(e, \delta) \wedge (\theta_{\text{hard}} < \delta)))$
Jim sings *better than* Mary.	$\exists e_1 \exists e_2(\text{sing}(e_1) \wedge (\text{subj}(e_1) = \text{jim}) \wedge \text{sing}(e_2) \wedge (\text{subj}(e_2) = \text{mary}) \wedge \exists \delta(\text{good}(e_1, \delta) \wedge \neg\text{good}(e_2, \delta)))$
Bob drives *as carefully as* John.	$\exists e_1 \exists e_2(\text{drive}(e_1) \wedge (\text{subj}(e_1) = \text{bob}) \wedge \text{drive}(e_2) \wedge (\text{subj}(e_2) = \text{john}) \wedge \forall \delta(\text{careful}(e_2, \delta) \rightarrow \text{careful}(e_1, \delta)))$

Table 1: Logical forms of adverbs and their comparative and equative forms

evaluate the system on various datasets to check whether the system can perform linguistically challenging inferences without compromising the accuracy of each basic theory.

More specifically, we build on the system presented in Haruta et al. (2020), which implements degree semantics for comparatives and generalized quantifiers. The system is limited, however, in that it does not implement event semantics, and hence does not handle inferences with adverbs and related constructions. Also, it only covers a small portion of the generalized quantifiers discussed in the linguistics literature and does not handle inferences requiring lexical knowledge, thus being confined to purely logical inferences. To test the compatibility of event semantics and degree semantics, we add a layer of event semantics and lexical knowledge to that system and attempt to broaden its empirical coverage consistently.

To evaluate the system, we assess its capacity to handle (i) a set of logical compositional inferences and (ii) their interaction with lexical knowledge. For (i), we use FraCaS (Cooper et al., 1994), which contains various semantically complex inferences and CAD (Haruta et al., 2020), which contains complex inferences with adjectives and comparatives. To our knowledge, there is no linguistically controlled dataset that contains inferences with adverbs, so we create a set of logical inferences with adverbs and related constructions and use it to evaluate the system. For (ii), we use MED (Yanaka et al., 2019), which contains problems with monotonicity inferences and lexical knowledge, and SICK (Marelli et al., 2014), which focuses on lexical inferences combined with linguistic phenomena such as negation and quantifiers. In addition, we use HANS (McCoy et al., 2019), which is designed to probe the capacity of NLI models based on deep learning (DL) and contains structural inferences, including those concerning adjectives and adverbs. The results show that our system achieves high accuracies across various datasets, supporting our claim that event semantics and degree semantics can be effectively combined for NLI.

2   System architecture

For the implementation of an NLI system[1], we follow the basic architecture of Haruta et al. (2020). Input sentences (i.e., a set of premises $P_1, \ldots, P_n$ and a hypothesis H) are mapped to CCG derivation trees using off-the-shelf CCG parsers. To accommodate output derivation trees in formal semantic analysis, various tree transformations are applied. Then, using a set of semantic templates that assign lambda-expressions to CCG categories (Steedman, 2000), the output trees are mapped to logical forms, which are formulas of first-order logic (FOL) with equality and arithmetic operations. This provides a set of formulas $P_1', \ldots, P_n'$ for the premises and H' for the hypothesis. Then, an FOL theorem prover tries to prove $P_1' \wedge \cdots \wedge P_n' \rightarrow H'$. If it is successful, the system outputs *yes* (entailment). Otherwise, the system tries to prove $P_1' \wedge \cdots \wedge P_n' \rightarrow \neg H'$ and outputs *no* (contradiction). If both attempts fail, the system outputs *unknown* (neutral). For our purposes, we add two components to this basic architecture: (i) we extend the set of the semantic template to accommodate event semantics and various types of generalized quantifiers, and (ii) we add a mechanism to insert lexical knowledge before theorem proving.

Combining event semantics and degree semantics   We use standard neo-Davidsonian event semantics (Parsons, 1990), which analyzes sentences as involving quantification over events. For instance, the sentence *John ran* is analyzed as $\exists e(\text{run}(e) \wedge (\text{subj}(e) = \text{john}))$, where subj is a function term that associates an event to its participant (subject). A sentence containing an adverb, for example, *John ran slowly*, is analyzed as $\exists e(\text{run}(e) \wedge (\text{subj}(e) = \text{john}) \wedge \text{slowly}(e))$, where the adverb *slowly* acts as a predicate of an event. This allows us to derive an inference from *John ran slowly* to *John ran*, that is, an inference to drop adverbial phrases.

---

[1] GitHub repository with code and data: `https://github.com/izumi-h/ccgcomp`

Type	Example	Logical form
upward	*Many* people cried.	$\exists x(\mathsf{people}(x) \wedge \mathsf{many}(x, \theta_{\mathsf{many}}(\mathsf{person})) \wedge \exists e(\mathsf{cry}(e) \wedge (\mathsf{subj}(e) = x)))$
downward	*Less* than five students laughed.	$\neg\exists x(\mathsf{student}(x) \wedge \mathsf{many}(x, 5) \wedge \exists e(\mathsf{laugh}(e) \wedge (\mathsf{subj}(e) = x)))$
non-monotone	*Exactly* eleven boys play soccer.	$\exists x(\mathsf{boy}(x) \wedge \mathsf{many}(x, 11) \wedge \exists e(\mathsf{play}(e) \wedge (\mathsf{subj}(e) = x) \wedge (\mathsf{obj}(e) = \mathsf{soccer})))$
		$\wedge \forall x \forall \delta(\mathsf{boy}(x) \wedge \mathsf{many}(x, \delta) \wedge \exists e(\mathsf{play}(e) \wedge (\mathsf{subj}(e) = x) \wedge (\mathsf{obj}(e) = \mathsf{soccer})) \rightarrow (\delta < 12))$

Table 2: Logical forms of upward, downward, and non-monotonic generalized quantifiers

Lexical relationship	antonym	hypernym	hyponym	synonym, similar, inflection, derivation
Axiom	$\forall x(F(x) \rightarrow \neg G(x))$	$\forall x(F(x) \rightarrow G(x))$	$\forall x(G(x) \rightarrow F(x))$	$\forall x(F(x) \leftrightarrow G(x))$

Table 3: Correspondence between the lexical relationships and the forms of inserted axioms

For degree semantics, we use the one presented in Haruta et al. (2020). It analyzes a gradable adjective *tall* as a binary predicate $\mathsf{tall}(x, \delta)$ holding of an entity x and a degree δ (Cresswell, 1976). Thus, the sentence *Chris is 5 feet tall* is mapped to the logical form $\mathsf{tall}(\mathsf{chris}, 5\ \mathsf{feet})$. When a degree expression like *5 feet* is absent, δ is set to a default value (for which we use a fixed constant); e.g., *Chris is tall* is mapped to $\mathsf{tall}(\mathsf{chris}, \theta_{\mathsf{tall}})$. Comparative expressions are analyzed in terms of first-order logic, using the so-called A-not-A analysis (Seuren, 1973; Klein, 1982; Schwarzschild, 2008); for example, *Chris is taller than Alex* is analyzed as $\exists\delta(\mathsf{tall}(\mathsf{chris}, \delta) \wedge \neg\mathsf{tall}(\mathsf{alex}, \delta))$, which asserts that there is a degree δ of tallness that Chris satisfies but Alex does not. See Haruta et al. (2020) for more detail. This analysis can be naturally extended to adverbial phrases and their comparative forms. Table 1 shows logical forms of basic constructions, where adverbs like *slowly* are treated as binary predicates of an event and a degree.

Monotonicity and generalized quantifiers   In degree semantics, generalized quantifiers such as *many* can be analyzed not as higher-order expressions (as in Montague's tradition) but as two-place predicates such as $\mathsf{many}(x, n)$, which reads "a composite entity x consists of at least n individuals" (Hackl, 2000; Rett, 2018). Haruta et al. (2020) implements this semantics in their system but does not deal with downward quantifiers (e.g., *few* and *less than five*) or non-monotonic quantifiers (e.g., *exactly five*). It is known that these two types of quantifiers pose a problem for the compositional treatment of quantifiers as predicates (van Benthem, 1986; Buccola and Spector, 2016). The problem is that downward and non-monotonic quantifiers need to take scope over the entire clause. For example, in the case of the downward quantifier *less than* in *Less than five students laughed*, the negation needs to take scope over the entire clause, as shown in Table 2. To solve this problem, we add syntactic features to CCG categories in the derivation trees in the post-processing process. Downward quantifiers such as *few* and *less than five* are assigned the category N/N (which is the same as an adjective like *tall*) in CCG parsers. We modify this to N_{down}/N, which triggers the derivation in which the negation takes the outermost scope. Similarly, we assign syntactic categories like N_{nm}/N to non-monotonic quantifiers such as *exactly* and *only*. This enables us to derive the desired logical forms as in Table 2.

Insertion of lexical knowledge   To test the compatibility of logical inferences and inferences involving lexical knowledge, we implement a mechanism to search for useful axioms drawn from knowledge bases before the process of theorem proving. The strategy is similar to the one used in previous studies (Martínez-Gómez et al., 2017) in which the system searches for lexical relations from WordNet (Miller, 1995) and VerbOcean (Chklovski and Pantel, 2004). More specifically, for each predicate F appearing in the set of formulas for given premises, if there is a predicate G appearing in the formula for the hypothesis such that (i) F and G have the same semantic type (e.g., the type of predicate of events) and (ii) F has a lexical relationship with G, then we add an axiom of the relevant form depending on the type of lexical relation. Following Martínez-Gómez et al. (2017), we use a total of seven relationships and add the corresponding axioms, as shown in Table 3.

3   Experiments

Experimental settings   We use three CCG parsers, namely, C&C (Clark and Curran, 2007), Easy-CCG (Lewis and Steedman, 2014), and depccg (Yoshikawa et al., 2017), for CCG parsing, and we use Tsurgeon (Levy and Andrew, 2006) for tree transformation. For CCG parsing and tree transformation,

Dataset	Label	ID	Example (premises and hypothesis)	Gold label
FraCaS	*Com*	241	P_1: ITEL won more orders than APCOM lost. P_2: APCOM lost ten orders. H: ITEL won at least eleven orders.	Yes
MED	*gq*	485	P_1: Exactly 12 aliens threw some tennis balls. H: Exactly 12 aliens threw some balls.	Unknown
	gqlex	176	P_1: Few aliens saw birds. H: Few aliens saw doves.	Yes
SICK	–	1357	P_1: A puppy is repeatedly rolling from side to side on its back. H: A dog is rolling from side to side.	Yes
HANS	*constituent*	23991	P_1: The actors contacted the president, or the lawyers recommended the managers. H: The lawyers recommended the managers.	Unknown
CAD+	*MA*	115	P_1: Exactly seven students smiled. H: At most nine students smiled.	Yes
		157	P_1: Ann runs as fast as Luis does. P_2: Ann runs slowly. H: Luis runs fast.	No

Table 4: Examples of entailment problems from the FraCaS, MED, SICK, HANS, and CAD+ datasets

FraCaS

Section	GQ	Adj	Com	Att
#All	73	22	31	13
Maj	.49	.41	.61	.62
RB	.73	.45	.52	.69
MN	.77	.68	.48	.77
LP	.93	.73	–	.92
HR	.95	.95	.84	–
Ours	.97	.82	**.90**	.92
+rule	**.99**	**.95**	**.90**	**.92**

MED

Label	gq	gqlex
#All	498	691
Maj	.58	.63
BERT	.56	.58
BERT+	.54	.68
RB	.57	.55
HR	.84	–
Ours	**.96**	**.92**

SICK

#All	4927
Maj	.57
RB	.56
LP	.81
MG	**.83**
Ours	.82

HANS

Gold	yes	unknown
#All	501	501
Maj	.50	.50
RB	**1.0**	.56
Ours	.97	**.78**

CAD+

Label	CAD	MA
#All	105	134
Maj	.48	.43
RB	.58	.59
HR	.77	–
Ours	**.83**	**.89**

Table 5: Accuracy on the FraCaS, MED, SICK, HANS, and CAD+ datasets

we use the same setting as in Haruta et al. (2020). We use a set of semantic templates for mapping CCG trees to logical forms. The templates are specified for 528 categories and 138 lemmas in total. The Tsurgeon script has 126 clauses for tree mapping. For POS tagging, we use the C&C POS tagger for C&C and spaCy[2] for EasyCCG and depccg. We use the FOL prover Vampire[3] for theorem proving.

Datasets For evaluation, we use five datasets. Table 4 shows some examples. (1) FraCaS has nine sections, of which we use four: Generalized Quantifiers (*GQ*), Adjectives (*Adj*), Comparatives (*Com*), and Attitudes (*Att*). (2) MED collects inferences with generalized quantifiers. We use a portion of the dataset taken from linguistics papers, which are divided into non-lexical inferences (*gq*: 498 problems) and inferences involving lexical knowledge (*gqlex*: 691 problems). (3) For SICK, we use the 2014 version of SemEval (Marelli et al., 2014). (4) For HANS, we randomly choose 1002 problems labelled as *constituent*, *lexical overlap*, and *subsequence* from the entire test set (30,000 problems), which are divided into entailment (*yes*) and non-entailment (*unknown*) problems. (5) CAD+ has two sections. CAD (Haruta et al., 2020) contains 105 inference problems concerning adjectives and comparatives, which are linguistically challenging but missing from FraCaS. We also create a new set of problems, called *MA* for monotonicity and adverbial phrases, which follows the patterns in CAD. It has 134 problems in total. Of these 134 problems, 69 are single-premise problems, and 65 are multi-premise problems. The distribution of gold answer labels is (*yes/no/unknown*) = (57/36/41).

3.1 Results and discussion

Table 5 shows the experimental results on all datasets. *Maj* is the accuracy of the majority baseline and *Ours* is the accuracy of our system. For FraCaS, *+rule* shows the accuracy achieved by the addition of hand-coded rules, which correct the errors in POS tagging and lemmatization, as described in Haruta et al. (2020). Given that MED and HANS use binary labels (*yes* and *unknown*), for these two datasets we modify the system so that it outputs *yes* if the hypothesis can be proved from the premise; otherwise, the output is *unknown*. We compare our system with previous logic-based systems and DL-based systems.

FraCaS We use three logic-based systems: MN (Mineshima et al., 2015), LP (Abzianidze, 2016), and HR (Haruta et al., 2020). These are systems based on CCG parsing and theorem proving. For a DL-based

system, we use a state-of-the-art model, RoBERTa (RB) (Liu et al., 2019), trained on MultiNLI (Williams et al., 2018), using the implementation provided in AllenNLP (Gardner et al., 2018). Our system achieved nearly 100% accuracy and outperformed the DL-system by a large margin. FraCaS-241 in Table 4 is a complex inference with numerical expressions; this problem is solved by our system but neither by the other logic-based systems nor by the DL-based system (RB). Our system also improved its predecessor (HR) in that it can handle inferences involving clausal comparatives (FraCaS 239–241).

MED For the results on MED in Table 5, BERT shows the performance of a BERT model fine-tuned with MultiNLI and BERT+ shows that of a BERT model with data augmentation for monotonicity inferences in addition to the MultiNLI training set. Both models were tested in Yanaka et al. (2019). Our system outperformed both the logic-based system (HR) and the DL-based systems. MED-176 and MED-485 in Table 4, which involve a downward quantifier (*few*) and a non-monotonic quantifier (*exactly 12*), respectively, are examples that our system correctly solved while the DL-models did not.

SICK For SICK, MG (Martínez-Gómez et al., 2017) is a system based on CCG parsing with compositional event semantics and theorem proving. Our system outperformed the DL-based system (RB) and achieved comparable results with the logic-based systems (LP and MG), showing that the combination of event semantics and degree semantics is compatible with the insertion of lexical knowledge. For example, SICK-1357 in Table 4 is an example involving the lexical inference from *puppy* to *dog*; our system correctly predicted the *yes* label for this problem, while the DL-based system (RB) predicted the *no* label.

HANS McCoy et al. (2019) reported that DL-based systems tend to erroneously output *yes* for cases in which the hypothesis was a constituent or a sub-string of the premise, such as disjunctive sentences (e.g., HANS-23991 in Table 4). To see how a system performs on these cases, we present the accuracy for each gold answer label (*yes* and *unknown*). While the accuracy when the gold label was *yes* was close to 100% in both our system and the DL-based system (RB), the accuracy of our system was higher than that of RB when the label is *unknown* (78% vs. 56%). One of the reasons for the relatively low accuracy (78%) of our system in comparison with the performance on the other datasets is parse error; HANS contains syntactically complex sentences such as *The actor paid in the library recognized the lawyers* (HANS-13628, *subsequence*), for which the current CCG parsers output incorrect parses. Another reasons is inference involving a modal adverb, e.g., the inference from *Probably the secretary admired the athlete* to *The secretary admired the athlete* (HANS-24034). The gold label is *unknown*, but our system predicts *yes*, because any adverb can be dropped in the current implementation. A more fine-grained classification of adverbs will be needed to handle this type of inference.

CAD+ For CAD+, our system outperformed the previous logic-based system (HR) and the DL-based system (RB). Our system was able to solve the inference involving numerical computation (CAD-115) and antonym conversion for adverbs (CAD-157) in Table 4, while RB incorrectly predicted *no* for CAD-115 and *yes* for CAD-157. However, some problems with adverbial expressions remain. For example, the sentence *Jones drives more carefully today than yesterday* (MA-183) conjoins two adverbs *today* and *yesterday* by *than*. The current system does not derive the correct logical form for this type of complex coordinate structure formed by *than*-clauses. A further improvement of CCG parsing would be needed to handle such complex coordinate constructions.

4 Conclusion

We have presented a logic-based NLI system that combines event semantics and degree semantics and evaluated the system on various datasets containing semantically challenging inferences. The results showed that the combination of event semantics and degree semantics is viable and works well on the type of complex logical inferences for which standard DL-based systems show poor performance. This study contributes to the study of computational modeling and the evaluation of formal semantic theories, as well as to the creation of challenging NLI problems that DL-based models need to address.

References

Lasha Abzianidze. 2016. Natural solution to FraCaS entailment problems. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 64–74. Association for Computational Linguistics.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 277–286. College Publications.

Brian Buccola and Benjamin Spector. 2016. Modified numerals and maximality. *Linguistics and Philosophy*, 39(3):151–199.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33–40. Association for Computational Linguistics.

Stephen Clark and James R Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. 1994. FraCaS–a framework for computational semantics. *Deliverable*, D6.

Max J Cresswell. 1976. The semantics of degree. In Barbara Partee, editor, *Montague Grammar*, pages 261–292. Academic Press.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press, Pittsburgh, PA.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6. Association for Computational Linguistics.

Martin Hackl. 2000. *Comparative Quantifiers*. Ph.D. thesis, Massachusetts Institute of Technology.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. Logical inferences with comparatives and generalized quantifiers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL): Student Research Workshop*, pages 263–270. Association for Computational Linguistics.

Christopher Kennedy. 1997. *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. Ph.D. thesis, University of California, Santa Cruz.

Ewan Klein. 1982. The interpretation of adjectival comparatives. *Journal of Linguistics*, 18(1):113–136.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2231–2234.

Mike Lewis and Mark Steedman. 2014. A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

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'14)*, pages 216–223.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2017. On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 1, Long Papers*, pages 710–720. Association for Computational Linguistics.

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 (ACL)*, pages 3428–3448. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2055–2061. Association for Computational Linguistics.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

Terence Parsons. 1990. *Events in the Semantics of English*. MIT press.

Jessica Rett. 2018. The semantics of *many*, *much*, *few*, and *little*. *Language and Linguistics Compass*, 12(1).

Roger Schwarzschild. 2008. The semantics of comparatives and other degree constructions. *Language and Linguistics Compass*, 2(2):308–331.

Pieter A. M. Seuren. 1973. The comparative. In F. Kiefer and N. Ruwet, editors, *Generative Grammar in Europe*, pages 528–564. Riedel, Dordrecht.

Arnim von Stechow. 1984. Comparing semantic theories of comparison. *Journal of Semantics*, 3(1-2):1–77.

Mark Steedman. 2000. *The Syntactic Process*. MIT press.

Johan van Benthem. 1986. *Essays in logical semantics*. Reidel.

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 (NAACL-HLT)*, pages 1112–1122. 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. Association for Computational Linguistics.

Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. 2017. A* CCG parsing with a supertag and dependency factored model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 277–287. Association for Computational Linguistics.