Curing the SICK and Other NLI Maladies

Aikaterini-Lida Kalouli*
Center for Information and Language Processing (CIS), LMU Munich
kalouli@cis.lmu.de

Hai Hu*
Shanghai Jiao Tong University
School of Foreign Languages
hu.hai@sjtu.edu.cn

Alexander F. Webb
Indiana University Bloomington
Department of Philosophy
afwebb@iu.edu

Lawrence S. Moss
Indiana University Bloomington
Department of Mathematics
lmoss@indiana.edu

Valeria de Paiva
Topos Institute
valeria.depaiva@gmail.com

Against the backdrop of the ever-improving Natural Language Inference (NLI) models, recent efforts have focused on the suitability of the current NLI datasets and on the feasibility of the NLI task as it is currently approached. Many of the recent studies have exposed the inherent human disagreements of the inference task and have proposed a shift from categorical labels to human subjective probability assessments, capturing human uncertainty. In this work, we show how neither the current task formulation nor the proposed uncertainty gradient are entirely suitable for solving the NLI challenges. Instead, we propose an ordered sense space annotation, which distinguishes between logical and common-sense inference. One end of the space captures non-sensical inferences, while the other end represents strictly logical scenarios. In the middle of the space, we find a continuum of common-sense, namely, the subjective and graded opinion

* Equal contributions. Corresponding authors: Aikaterini-Lida Kalouli, Hai Hu.

Action Editor: Kevin Duh. Submission received: 23 November 2021; revised version received: 20 September 2022; accepted for publication: 26 September 2022.

https://doi.org/10.1162/coli_a_00465

of a "person on the street." To arrive at the proposed annotation scheme, we perform a careful investigation of the SICK corpus and we create a taxonomy of annotation issues and guidelines. We re-annotate the corpus with the proposed annotation scheme, utilizing four symbolic inference systems, and then perform a thorough evaluation of the scheme by fine-tuning and testing commonly used pre-trained language models on the re-annotated SICK within various settings. We also pioneer a crowd annotation of a small portion of the MultiNLI corpus, showcasing that it is possible to adapt our scheme for annotation by non-experts on another NLI corpus. Our work shows the efficiency and benefits of the proposed mechanism and opens the way for a careful NLI task refinement.

1. Introduction

In recent years, Natural Language Inference (NLI) has emerged as one of the most prominent tasks for evaluating a system's or a model's ability to perform Natural Language Understanding and reasoning. Typically, the NLI task consists of finding whether a sentence (the premise P) entails, contradicts, or is neutral with respect to another sentence (the hypothesis H) (MacCartney 2009). A surge of interest has resulted in the creation of huge datasets and the training of massive deep models. The trained models have been shown to achieve state-of-the-art results, sometimes even outperforming what is considered human performance (Liu et al. 2019a; Pilault, Elhattami, and Pal 2020). However, it has also been repeatedly shown that such models learn to capitalize on superficial cues and fail to generalize to more complex reasoning patterns. On the one hand, this can be attributed to the current NLI datasets themselves: They contain biases and artifacts that are mostly responsible for the high performance of the models, and their absence from other datasets naturally leads to worse models (Poliak et al. 2018; Gururangan et al. 2018; Geva, Goldberg, and Berant 2019). On the other hand, the generalization difficulties of such models also stem from the simplicity of the current datasets, which does not allow the models to learn to efficiently solve inferences requiring true (semantic) understanding of language (Dasgupta et al. 2018; Nie, Wang, and Bansal 2018; Naik et al. 2018; Glockner, Shwartz, and Goldberg 2018; McCoy, Pavlick, and Linzen 2019; Richardson et al. 2020; Yanaka et al. 2020). Against this backdrop, recent efforts have focused on the reliability and suitability of current datasets and on the feasibility of the NLI task as it is currently approached. One strand of research has discovered opposing annotations in datasets and has attempted to either correct these errors to make the corpora more reliable for training and testing models (e.g., Kalouli, Real, and de Paiva 2017a,b, 2018; Hu, Chen, and Moss 2019), or convert the 3-label task to a 2-label task to avoid some of the controversies (entailment vs. non-entailment; e.g., McCoy, Pavlick, and Linzen 2019). Another strand of research (Pavlick and Kwiatkowski 2019; Nie, Zhou, and Bansal 2020; Chen et al. 2020) has argued that the current NLI task formulation is unsuitable to capture the inherent disagreements and uncertainty involved in human reasoning. This series of work has proposed a refinement of the NLI task to be able to capture more subtle distinctions in meaning, by shifting away from categorical labels to human subjective probability assessments of a specific scenario, based on an ordinal or graded scale of "certainty."

These observations around inference are not novel, though. Already Glickman, Dagan, and Koppel (2005), Versley (2008), Poesio and Artstein (2005), and Pavlick and Callison-Burch (2016) have argued that humans' inferences are uncertain, context-dependent, and have a probabilistic nature, while in previous work from some in our group (Kalouli et al. 2019) we have proposed to "start devising corpora based on the

notion of human inference which includes some inherent variability." Thus, in this work we set out to shed light on the aspects of NLI that make it so error-prone and uncertain by conducting an extensive study of the SICK corpus (Marelli et al. 2014b). We find that it is not necessarily uncertainty that triggers different, sometimes even opposing, annotations, but rather a confrontation: *common-sense vs. logical inference*. This distinction is not novel: It dates back at least to Zaenen, Karttunen, and Crouch (2005), Manning (2006), and Crouch, Karttunen, and Zaenen (2006), but it has never been explored to its full potential. We show that inference cannot be entirely and efficiently modeled either as a single-label task or as a graded, probability distribution task. The NLI label space should best be seen as an *ordered sense space*. The one end of the ordered space features non-sensical inferences, namely, invalid inferences that could not hold for a given pair (e.g., *P: The man is shouting. H: The man is quiet* could never be an entailment, assuming the same man at the same moment), while the other end represents strictly logical inferences that a judge (or a logician, for that matter) would take into account (e.g., *P: The woman is cutting a tomato. H: The woman is slicing a tomato* is a neutral pair and not an entailment, because strictly speaking the woman does not need to be slicing the tomato; maybe she is cutting squares.) The two ends are solid points in the space and can thus be characterized by a single, distinct label, much as it is done within the traditional NLI task. In the middle of the space, we find common-sense, that is, the opinion of a "person on the street," which is, however, subjective, graded and is thus represented by a continuum (e.g., for the common-sense of most people, the previous example of *cutting/slicing* is probably an entailment, but some might still think that this is neutral, even based on their common-sense). Thus, in contrast to the ends of the ordered space, the in-between common-sense continuum captures a distribution of potentially different labels. Each inference label can be placed on a different area of the ordered sense space. We show how this inference scheme can capture the challenges posed by NLI data more efficiently than the traditional single-label approach and the more recent certainty gradient, and we open the way for the creation of high-quality, reliable NLI datasets.

To achieve an extensive study of the SICK corpus and to be able to propose the ordered sense space annotation, it is necessary to pursue a series of steps. First, we take up and reconsider our previous efforts to correct the corpus (Kalouli, Real, and de Paiva 2017a,b, 2018; Hu, Chen, and Moss 2019). These previous attempts were mainly manual, concentrated on parts of the SICK corpus, and were done by "experts," that is, researchers conversant with the notion of inference as it appears in academic areas like semantics and logic. This study is meant to investigate the *entire* corpus, and thus we exploit the power and performance of several symbolic NLI systems that were introduced at the same time as this work. Specifically, we are able to use symbolic systems to automatically and reliably annotate large parts of the corpus. We attempt to hand-annotate the remaining corpus based on detailed guidelines, using concepts from semantics such as coreference, aspect, and vagueness, but keeping the original single-label NLI task formulation. The failure of this attempt points to the necessity for a new annotation approach, offered through the sense space. The efficiency of the proposed method is proven by detailing how it affects performance of neural models under various training and testing settings.

Thus, in this paper we make the following contributions: (1) we re-annotate the entire SICK corpus and thus provide a high-quality dataset of NLI that can be used for reliable training and testing, (2) we shed light on the issues that plague NLI datasets like SICK, (3) we show that it is possible to develop clear guidelines for the NLI task, but no degree of specificity of guidelines will completely address the inherent human

disagreements, (4) to this end, we propose the ordered sense space annotation, which distinguishes between what logicians would annotate in a precise and quantifiable sense from what "people on the street" might say, and we show why this annotation is more suitable than a single-label method or a certainty gradient, (5) we show marked differences in the performance of neural models when tested on the dataset annotated with the proposed sense annotation, and (6) we use a small crowd annotation of 100 examples from the MultiNLI corpus (sampled from chaosNLI) to demonstrate the feasibility of scaling up our proposed scheme for non-expert annotators on other NLI corpora. The refined SICK corpus and the crowd annotations of the 100 examples from MultiNLI using our proposed scheme can be found at `https://github.com/huhailinguist/curing-SICK`.

2. Related Work

2.1 Existing Annotation Schemes

In formal semantics, a common definition of **entailment** (Chierchia and McConnell-Ginet 2001) determines that P entails H if H is true in every possible world in which P is true. Similarly, **contradiction** is defined as the conjunction of a proposition and its denial, that is, P and H are contradictory if there is no possible world in which P and H are both true (Mates 1972). However, both these definitions are too strict for real-world scenarios because they require that H is true in *all* worlds where P is true, and it is usually not possible to know if something is true in all possible worlds to begin with. Thus, Dagan, Glickman, and Magnini (2005) offer an informal, more practical definition as part of the first RTE challenge:

> *"p entails h if, typically, a human reading p would infer that h is most likely true. . . [assuming] common human understanding of language [and] common background knowledge"*

The definition was supposed to be preliminary and invoke further discussions and refinements based on the needs of the task. Indeed, the discussion was initiated by Zaenen, Karttunen, and Crouch (2005), who argue for the need to make the task more precise by circumscribing whether and what type of world-knowledge is required to solve a specific inference, and to explicitly annotate for the different types of inference required, for example, entailment, conventional and conversational implicatures, and so forth. Manning (2006), on the other hand, argues for a more "natural" task, where untrained annotators can label pairs in real-life settings, without guidelines, definitions, or further explicit annotation information. Since then, there has been no formal consensus on whether a precise, but possibly distant from daily life, inference or a "real-life," but possibly inconsistent, annotation approach is more suitable. However, there has been an informal preference for the latter approach, as is evident in the huge datasets promoting natural inferences over rigorous annotation processes (Bowman et al. 2015; Williams, Nangia, and Bowman 2018).

The first such large dataset was the SICK corpus (Marelli et al. 2014b), which uses a 3-way annotation scheme, moving away from the entailment vs. non-entailment task into the more general inference task. The scheme utilizes entailment, contradiction, and neutral relations. Annotators are asked to judge the pairs as naturally as possible. The annotators are not experts but rather crowdsourcing workers. They are not given information on the source of the sentences nor specific guidelines to adhere to, but rather a short example for each inference relation. The annotation scheme pursued

in the later, much larger datasets SNLI (Bowman et al. 2015) and MNLI (Williams, Nangia, and Bowman 2018) is very similar. For these datasets annotators are asked to judge whether a given sentence can be a definitely true or a definitely false description of another sentence (entailment and contradiction relations, respectively) or whether none of these two applies (neutral relation). Again, strong emphasis is given to the naturalness of the annotation process and on the annotators being lay people. This 3-way annotation scheme had already been used for the smaller but also very complex earlier datasets of the RTE challenge. Specifically, RTE-4 (Giampiccolo et al. 2008), RTE-5 (Bentivogli et al. 2009), RTE-6 (Bentivogli et al. 2010), and RTE-7 (Bentivogli et al. 2011) were constructed to include all 3 inference labels. In contrast, in the first three RTE challenges (RTE-1, Dagan, Glickman, and Magnini [2005]; RTE-2, Bar-Haim et al. [2006]; and RTE-3, Giampiccolo et al. [2007]), a 2-way scheme was used, with a label allotted for entailment and a label allotted for non-entailment, following Dagan, Glickman, and Magnini's (2005) definition presented above.

The 2-label scheme, however, is not only an annotation practice used in earlier datasets; recent research has also opted for this annotation scheme. For example, the 2-label annotation is used in the HELP dataset (Yanaka et al. 2019b), which is designed to improve the performance of neural models specifically for monotonicity reasoning. The HELP scheme provides one label for entailment and one label for neutral. This differs from the 2-way classification schemes in the first three RTE challenge datasets merely in terms of presentation, since non-entailment indicates as much as neutral; neither distinguishes between a lack of entailment and a proper contradiction. To presumably overcome this limitation, in a pilot task of the RTE-3 challenge, the third category of *unknown* was added; yet the categories entailment, non-entailment, and unknown do not map neatly onto a 3-way scheme of entailment, contradiction, and neutral. "Unknown" signals, at best, an uncertainty regarding what sort of relationship holds between the pair of sentences. Only with a category dedicated to contradictions do we get a full, classical tripartite division of the space of possible relations between the pair. Despite this drawback of the 2-way classification, further efforts opt for it as they consider it an "easier" and "clearer" task. For example, the Heuristic Analysis for NLI Systems (HANS) dataset (McCoy, Pavlick, and Linzen 2019), an adversarial NLI corpus, uses a 2-way scheme of entailment and non-entailment and this is a conscious decision on the authors' part, as "the distinction between contradiction and neutral is often unclear" for the cases they wish to consider. Similarly, the SciTail dataset (Khot, Sabharwal, and Clark 2018) uses a 2-way scheme of entailment and neutral. The Monotonicity Entailment Dataset (MED) dataset (Yanaka et al. 2019a) also follows the same pattern in principle. It includes the third category of *unnatural*, but this is reserved for cases where either the premise or the hypothesis is ungrammatical or does not make sense. Thus, this category only screens out grammatical or other infelicities and remains at its core a 2-way annotation scheme. Although slightly different, the same 2-way classification goal is achieved by Bhagavatula et al. (2020) in their αNLI dataset. This dataset is developed to capture abductive reasoning. The two associated tasks, the Abductive Natural Language Inference (αNLI) task and the Abductive Natural Language Generation (αNLG) task, are not designed to track entailment. Rather, the former task involves taking a pair of observations and deciding between a pair of hypotheses which better explains the observations, assigning a label of plausible and implausible (or less plausible) to each hypothesis. The latter task involves generating a valid hypothesis for a pair of observations. Again, this 2-way classification scheme of plausible or implausible fails to capture finer-grained distinctions within the realm of abductive inference, but, on the other hand, it is easier for annotators to get right.

The classic 2-way and 3-way annotation schemes raise worries concerning their suitability. The 3-way scheme is more fine-grained and can efficiently capture human reasoning, which includes the notion of contradiction. However, it also leads to more error-prone and inconsistent annotations because often no clear line can be drawn between neutrality and contradiction. The 2-way scheme avoids this controversy, but cannot efficiently capture the whole realm of human inference, as it conflates consistent sentences with contradictory ones. Additionally, it has been shown that neither scheme is suitable for capturing the inherent human disagreements that occur in the inference task (Pavlick and Kwiatkowski 2019). An interesting experiment is performed by Nie, Zhou, and Bansal (2020) in their ChaosNLI dataset. The researchers create their dataset by collecting annotations for over 4,000 sentence pairs from SNLI, MNLI, and αNLI and aim to find the distribution of annotator agreement and disagreement in the population. They show that a significant number of the original majority labels fail to capture the new majority opinion among annotators, and that for a subset of sentence pairs large-scale human disagreement persists. Thus, recently there have been significant departures from the classical 2-way and 3-way schemes. For example, attempts have been made to use ordinal or even real-valued scales to capture likelihood and inherent human disagreement.

Zhang et al. (2017) develop their ordinal annotation scheme on the premise that common-sense inference is "possibilistic," in the sense that common-sense inferences do not hold with the sort of necessity accompanying logical entailment. For the annotation, they use the 5-point Likert scale of likelihood (very likely, likely, plausible, technically possible, impossible), based on the (Horn 1989) conception of scales of epistemic modality. Annotators are provided with one context sentence and one hypothesis sentence and then asked to rank the plausibility that the hypothesis is true "during or shortly after" the context, since "without this constraint, most sentences are technically plausible in some imaginary world." An open question about this proposed scale is why it countenances the *impossible* while limiting the opposite extremum to *very likely*. Though many inferences may only carry high confidence, many others certainly provide genuine entailments. In that case, one would expect a corresponding category of certainty, rather than very likely. Pavlick and Kwiatkowski (2019) experiment with a scaled rather than an ordinal annotation. They use a sliding scale from -50 to 50, with -50 indicating that the hypothesis is "definitely not true" given the premise, 50 indicating that the hypothesis is "definitely true" given the premise, and 0 indicating that the hypothesis is at least consistent with the premise. Raters are presented with a sliding bar and cannot view numbers on the slider. With their experimental results, the researchers show that NLI labels cannot efficiently be captured by a single aggregate score and that models trained to predict such a score do not learn human-like models of uncertainty. Instead, they propose that models learn a distribution over the labels—this proposal has parallels to our proposed annotation, as becomes clear in Section 5. Similar is the approach taken by Chen et al. (2020). With their Uncertain Natural Language Inference (UNLI) corpus, the researchers also make the move to subjective credence predictions with a refinement of the NLI corpus classification scheme using a real-valued scale. One key difference between the scheme utilized by Chen et al. (2020) and that of Pavlick and Kwiatkowski (2019) is that the former does not reduce the scale to a categorical classification.

2.2 The SICK Corpus

The SICK (Sentences Involving Compositional Knowledge) corpus (Marelli et al. 2014b) is an English corpus of 9,927 pairs, initially created to provide a benchmark

for compositional extensions of Distributional Semantic Models. The SICK corpus was created from captions of pictures with text mentioning everyday events and activities and non-abstract entities. The creation was a 3-step process. First, each caption was normalized to (a) exclude hard linguistic phenomena, such as modals, named entities, and temporal phenomena; (b) limit the amount of encyclopedic world-knowledge needed to interpret the sentences; and (c) make sure that complete sentences with finite verbs were included. Then, each normalized sentence was used to generate three new sentences based on a set of rules, such as adding passive or active voice and adding negations. A native speaker checked the generated sentences for grammaticality and removed ungrammatical ones. Each sentence was then paired with all three generated sentences. Each pair of sentences was annotated by five crowdsourcing workers for its inference relation (entailment, contradiction, neutrality) and its semantic similarity. We do not deal with semantic similarity in this paper. The inference relation was annotated in both directions, that is, annotators described the relation of sentence A with respect to sentence B and the relation of sentence B with respect to sentence A. Then, each pair was given a final label based on the judgment of the majority of the annotators.

At this point, the question of why we chose to perform our extensive NLI corpus analysis on the SICK corpus may arise. After all, the corpus is relatively small in comparison to other contemporary NLI datasets and it is simplified in many respects. But exactly these characteristics make it suitable for our purposes. First, SICK's size is ideal for our manual inspection and improvement. Alhough superseded by much larger datasets like SNLI (Bowman et al. 2015) and MNLI (Williams, Nangia, and Bowman 2018), SICK is small enough to be feasible for manual investigation and re-annotation. At the same time, it contains similar kinds of simplified data and language as the large datasets, for example, it does not include modals or implicatives, and thus it is representative enough of the kinds of inference included in larger state-of-the-art datasets (while allowing their manual detection and correction). Additionally, it might seem that SICK suffers from different issues than large datasets and that thus our findings are not transferable. The large datasets were shown to include annotation artifacts due to their collection process (i.e., people asked to create entailing, neutral, and contradicting sentences and thus looking for cognitive shortcuts to perform the task faster), which is different from the SICK process. However, Poliak et al. (2018) show that such artifacts are found both in human-judged datasets like SICK, where humans only label the inference relation, and in human-elicited datasets like SNLI and MultiNLI, where humans produce the hypothesis for a given premise. This is also discussed in Kalouli (2021). Thus, the findings of this study are transferable. Furthermore, SICK is uniform enough to teach us lessons and highlight challenges of annotation and inference. Also, SICK is chosen because it has already been studied before (cf. Kalouli, Real, and de Paiva [2017a,b, 2018]; Hu, Chen, and Moss [2019]) and has been shown to have annotation inconsistencies. Overall, using SICK should serve as a proof of concept for our approach and open the way for applying our proposed method to further corpora such as SNLI and MNLI or newly created corpora.

2.3 Correction Efforts for SICK

With the aim of using SICK as a training and testing corpus for automatic NLI systems, we set out to investigate its annotations. The investigation showed that the human judgments used in the SICK corpus can be erroneous, in this way lessening its usefulness for training and testing purposes. Thus, in our previous work, we tried to address

some of the issues, many of which had already been noted in the original paper of Marelli et al. (2014b).

First, we found that single entailments (i.e., A entails B, but B is neutral to A), were often confused with double entailments (i.e., bidirectional entailments: A entails B and B entails A), but such a confusion was more or less expected.[1] More surprising was our finding that given the way the corpus was annotated, contradictions were not symmetric. The work in Kalouli, Real, and de Paiva (2017a) re-annotated all single entailments of the corpus, creating a taxonomy of the most common errors. Further work in Kalouli, Real, and de Paiva (2017b) concentrated on manually investigating (and correcting) the contradiction asymmetries of the SICK corpus. Most such asymmetries are caused by the presence of indefinite articles in the sentences and the lack of a specific guideline about them.[2] So, this investigation was a harder endeavor: Because a contradiction requires that the two sentences refer to the same events and/or entities (Zaenen, Karttunen, and Crouch 2005; de Marneffe, Rafferty, and Manning 2008), we had to make sure that the referents can be made explicit and coreferring.[3] Thus, we had to make a semantic compromise and establish a guideline: assume that the pairs are talking about the same event and entities, no matter what definite or indefinite determiners are involved.[4] This guideline was necessary to allow us to correct the asymmetric contradictions, but, as shown in Kalouli et al. (2019), not sufficient to mitigate all issues of the dataset. The small-scale but controlled experiments performed in Kalouli et al. (2019, 2021) shed light on how human annotators often have opposing perspectives when annotating for inference, but still follow their own internal coherent perspective. Particularly, that work showed that the distinction between contradiction and neutrality is surprisingly hard for annotators, which makes it one of the main sources of inconsistent datasets. Thus, the work proposed to enhance the NLI task with two kinds of annotations: explanations for the specific decision of an annotator and a difficulty score of the annotation. With these additional annotations, one can more easily check whether the corpus created adheres to the given guidelines and one can also decide to treat differently inference instances that are either easier or more difficult.

Hu et al. (2020) reported another effort to partially correct SICK, which builds directly on Kalouli, Real, and de Paiva (2017a,b). In this study, the authors manually examined all the pairs where Kalouli, Real, and de Paiva (2017a,b) do not agree with the original SICK—about 400 pairs—and found that there are issues with roughly 100 of them. This suggests that disagreement in NLI annotation is probably far more common than many have previously assumed. This is also echoed by Pavlick and Kwiatkowski (2019), as well as Nie, Zhou, and Bansal (2020), who show that for SNLI, if one takes the majority label from 100 annotators instead of 5, the majority label would be different in 30% of the examples.

2.4 Symbolic Inference Systems

A number of symbolic NLI systems have been introduced or extended recently, showing high accuracy even for hard linguistic phenomena. We exploit their power and high performance to automatically and reliably re-annotate parts of the corpus SICK.

1 The confusion being that single entailments were still annotated as being double entailments.
2 Marelli et al. (2014b) already point out this weakness of their guidelines.
3 This issue was already discussed for RTE; see de Marneffe, Rafferty, and Manning (2008).
4 There are certain restrictions, though; see Kalouli, Real, and de Paiva (2017b) for more details.

Specifically, we use two different versions of the *ccg2lambda* system by Yanaka et al. (2018), the *LangPro* system by Abzianidze (2017), *MonaLog* by Hu et al. (2020), and *GKR4NLI* by Kalouli, Crouch, and de Paiva (2020) and Kalouli (2021).

ccg2lambda1 and ccg2lambda2. In the system presented by Yanaka et al. (2018), the inference pair is mapped to logical formulas based on neo-Davidsonian event semantics. More precisely, the work in Yanaka et al. (2018) uses *ccg2lambda*, a system that parses sentences into CCG (Combinatory Categorial Grammar) representations and then converts them into semantic representations. The researchers experiment with several CCG parsers, which lead to slightly different results. Here, we use two of the best settings reported, with the C&C parser (Clark and Curran 2007) (*ccg2lambda1*) and the EasyCCG (Lewis and Steedman 2014) parser (*ccg2lambda2*). The semantic representations produced are represented as directed acyclic graphs. Then, variables from the premise and the hypothesis are unified using a theorem proving mechanism and inference is performed based on Natural Deduction rules. The researchers also propose a method to perform phrase-to-phrase alignment to arrive at the inference relation between phrases. The best setting of the ccg2lambda system achieves the state-of-the-art results on SICK for logic-based models (84.3% in accuracy).

LangPro. The LangPro system (Abzianidze 2014, 2015, 2016, 2017) is also based on CCG and λ terms, but it uses an analytic tableau method and rules of Natural Logic (van Benthem 2008; Sánchez-Valencia 1991; MacCartney 2009) for proving. The CCG parser first produces a CCG parse tree, which is then translated to λ logical forms by the Lambda Logical Forms generator (LLFgen). An LLF-aligner is optionally used to align identical chunks of LLFs in the premise and hypothesis, so that these chunks can be considered as a whole without any internal structure. Finally, a theorem prover based on a first-order logic prover (Fitting 1990) returns the inference prediction. The rule inventory of the prover contains roughly 50 rules, which are manually coded (see Abzianidze [2014] for details of the rules). LangPro achieves 82.1% accuracy on SICK (Abzianidze 2016).

MonaLog. The MonaLog system is a symbolic system that makes use of monotonicity facts of quantifiers and other words for inference (Hu, Chen, and Moss 2019; Hu et al. 2020). It involves three major steps: (1) Polarity/Arrow tagging; (2) Generation based on knowledge base $\mathcal{K}$; and (3) Search. Specifically, given a premise text, MonaLog first tags all the tokens in the premise with polarity annotations ($\uparrow, \downarrow, =$), using the polarity annotation tool ccg2mono (Hu and Moss 2018). The *surface-level* annotations, in turn, are associated with a set of inference rules based on Natural Logic. These rules provide instructions for how to generate entailments and contradictions by span replacements over these arrows (which relies on a library of span replacement rules). A `generation and search` procedure is then applied to see if the hypothesis text can be generated from the premise using these inference rules. A *proof* in this model is finally a particular sequence of edits that derive the hypothesis text from the premise text rules and yields an entailment or contradiction. MonaLog's accuracy on SICK is 81.66%.

GKR4NLI. GKR4NLI (Kalouli, Crouch, and de Paiva 2020; Kalouli 2021) is inspired by the Entailment Contradiction Detection (ECD) algorithm (Bobrow et al. 2007; Crouch and King 2007) and uses the Natural Logic (van Benthem 2008; Sánchez-Valencia 1991; MacCartney 2009) style of inference. The system first converts the sentences of an inference pair into their GKR representations (Kalouli and Crouch 2018), that is, to semantic

graphs capturing conceptual, contextual, lexical, and morphosyntactic constraints of the sentence, and then it aligns matching terms across the sentences. Based on the specificities and the instantiabilities of the matched terms, the inference relation is computed. GKR4NLI achieves state-of-the-art results across datasets of different complexity (see Kalouli, Crouch, and de Paiva [2020] for more details); on the SICK test set accuracy reaches 78.5%.

3. First-round Re-annotation of SICK

Our previous efforts to manually correct parts of the SICK corpus were successful. However, they were tedious and extremely time-consuming. Thus, in this new endeavor to annotate the *entire* corpus, we try to get as many annotations as possible for "free" from the automated systems. In other words, we experiment with available tools and systems to automatically annotate parts of the corpus, without sacrificing precision and reliability. To this end, the available symbolic inference systems are put into use.

3.1 Automatic Annotation

A common characteristic of the symbolic inference systems presented is that they have high precision for entailing and contradictory pairs. Because they are logic-based systems, they always have to rely on some hard "evidence" or proof (e.g., some WordNet sense) to predict entailment or contradiction. Thus, their precision for these labels is very high. On the other hand, the neutral label is not as reliable: Most of these systems fall back to the neutral relation whenever they cannot find evidence for an entailment or a contradiction. This does not mean, however, that the pair is necessarily a neutral pair; the evidence might not be codified. Given this behavior of the symbolic systems, it is necessary to make a distinction on how we consider and how much we trust neutral and non-neutral predictions. Therefore, after running the entire SICK corpus through each of the five symbolic systems, the quality and reliability of the predicted labels is judged separately for the non-neutral (entailment and contradiction) and neutral pairs. For both categories we also use the original SICK label as an additional indication of the correct label.

For the non-neutral predictions, only pairs for which the majority of the symbolic systems (at least three) agree on the label are taken as reliably annotated. If three symbolic systems have found some proof for an entailment or a contradiction, there is a very strong chance that this is indeed true; the other two systems probably missed this information.[5] But as we aim for a rigorous evaluation, we do not consider this criterion alone as adequate to accept the pair as reliably annotated. Instead, we also consider the original SICK label of the pair. If this does indeed agree with the majority predicted label of the symbolic systems, then we consider it correct. If not, then the pair has to be manually annotated. Figure 1 shows the decision-tree-style process we followed for the automatic annotation of the non-neutral pairs. Out of the 2,579 pairs for which at least three symbolic systems predict the same label, we are able to accept a total of 2,537, in which the predicted label also agrees with the SICK human label. There was no significant difference between the number of accepted entailments (E)

5 Since symbolic systems are based on the detection of a specific "hard" proof for their decision, the lack of such a proof will lead to a neutral and thus possibly a false label, disagreeing with the original one.

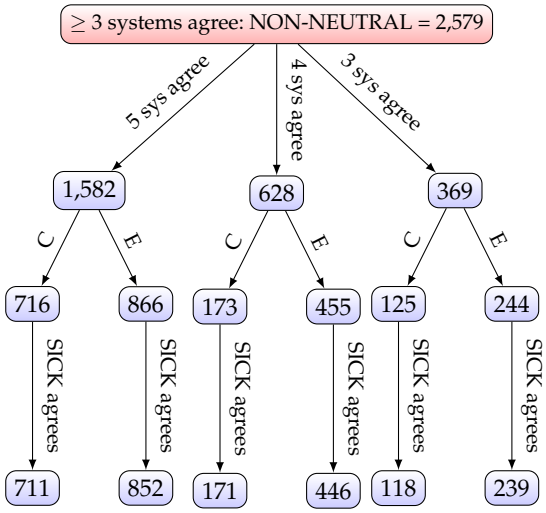

Figure 1
Non-neutral predicted labels of the five symbolic systems and the decision-tree-style process we follow for accepting them as reliable (or not). The figure shows a detailed break-down in each of the inference labels.

and contradictions (C) (98.2% and 98.6% were accepted, respectively). The remaining 42 pairs are collected for the manual annotation process (cf. Section 3.2).

For the neutral pairs we cannot allow the same decision process because the predicted labels are not as trustworthy. Therefore, we only consider reliable the pairs for which all 5 systems predict neutral N and the SICK original label is also N. All the rest of the pairs are collected to be manually checked. With this, the decision-tree-style process is formulated as shown in Figure 2. We see that 4,274 of the 5,344 pairs can be accepted as reliably annotated. For the other 1,070 pairs the SICK original label is either E or C and therefore a closer look is necessary.

After this automatic process, a total of 3,138 pairs (31.6% of the entire corpus) has to be manually annotated. Note that this automatic annotation can also be used on other corpora such as SNLI and MNLI; a main strength of symbolic systems is that they are not fine-tuned for a specific dataset and can thus deal with all kinds of data. Depending on the performance of the systems on other corpora the method of accepting some annotated pairs as reliable or not might need to be adjusted.

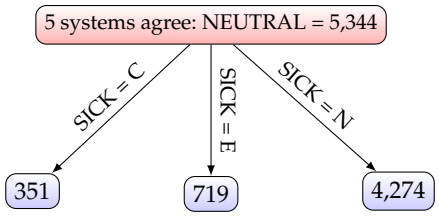

Figure 2
Neutral predicted labels of the five symbolic systems and the corresponding original label of SICK. Only the pairs for which the original SICK also delivers the N label are accepted as reliable.

3.2 Manual Annotation

Although the automatic annotation helped us speed up the process, the manual annotation was the product of two years' work. The manual annotation was performed by the authors, two of whom are English native speakers, while the other three have native-like English proficiency. Considering the error-prone nature of the task, which is also one of the main points of investigation in this work, we deem it absolutely necessary to adhere to common guidelines. For this we are able to capitalize on our experience from previous annotation efforts (see Section 2.3) and we attempt to consolidate serious disagreements. To this end, we first annotate a calibration set of 345 pairs. Each pair for which at least one annotator gives a different label is extensively discussed by the team and, if necessary, an appropriate new guideline is created.

3.2.1 Initial Guidelines. Based on our previous correction efforts and our calibration set discussions, we are able to put together a list of 15 guideline issues, presented in the following. These guidelines already foreshadow the challenges of the task and why an ordered sense space annotation is necessary.

Ungrammatical Pairs. We find pairs in which there is at least one grammatical error. In these cases, the pair is marked accordingly and is not annotated. Example: *P: A man in a car is pulling up beside a woman who is walking along a road. H: A man in a car is not pulling up beside a woman who is walking along a road.* We avoid annotation because each annotator might be performing a different "mental fix" and thus, the resulting annotation cannot be guaranteed to be consistent (cf. Kalouli, Real, and de Paiva 2017a,b). We find 71 ungrammatical pairs (0.7% of the entire corpus).[6]

Semantically Absurd Pairs. Apart from the ungrammatical pairs, we also find pairs in which one of the sentences is semantically absurd. As before, the pair is marked accordingly, and it is not annotated. Example: *P: A motorcycle rider is standing up on the seat of a white motorcycle. H: A motorcycle is riding standing up on the seat of the vehicle*. This treatment is necessary for the same "mental fix" reason mentioned above. We only find 41 such pairs (0.4% of the entire corpus).

Aspect/Tense. There are pairs where the tense and aspect of the predicates might influence the annotation, if taken literally. Example: *P: The adults and children **are** not **gathered** near an outdoor seating arrangement. H: The adults and children **are being gathered** near an outdoor seating arrangement.* In such cases, we choose to ignore such verbal properties and consider predicates as referring to the present and the non-progressive. Ignoring verbal properties is a common practice in such annotation projects (Cooper, Chatzikyriakidis, and Dobnik 2016; de Marneffe, Rafferty, and Manning 2008). In fact, during the creation of SICK there was an effort to eliminate such effects, but apparently not every instance could be detected.

Coreference. It is not trivial to decide when entities or events should be taken to refer to the same entity or event. Especially in SICK, this is vexing because the original

6 Note that for the later evaluation, all pairs with no labels default to the neutral label.

annotation was done as if the P and H were captions of a picture: the premise cannot contain everything in the picture, and so the objects referred to in it might or might not include everything of relevance. So, we first need to decide whether the subject entity and the verb event can be taken to refer to the same entity/event. If so, we can consider the sentences *coreferent* and annotate them accordingly; otherwise, elements cannot be made coreferent and the label should be N. Note that only coreferring entities or events can be taken to entail or contradict each other (Zaenen, Karttunen, and Crouch 2005; de Marneffe, Rafferty, and Manning 2008). Example: *P: A woman in red shirt is walking. H: A woman is moving.* Here, both the subject and the verb can be taken to refer to the same entity and event. However, there are also corner cases. Consider the pairs *P: A woman is walking. H: A man is walking* and *P: A woman is walking. H: A woman is dancing*. In the first example, the subjects are antonyms but the verbs corefer; we would not want to say that such a pair is neutral. So, we use the guideline that, if the verb is the same and the subjects are alternative antonyms (e.g., *man* vs. *woman*), then we can consider them coreferent and label such pairs as contradictory. In the second example, the subjects corefer but it is not clear what the verbs do. So, again, our agreed guideline says that if it is highly unlikely that the two events happen at the same time, then we should annotate as C, otherwise as N. In this example, it is likely that *walking* and *dancing* happen at the same time, since one might be walking during the dance. But in an example like *P: The man is fasting. H: The man is eating*, it is highly unlikely that the two events happen at the same time. Note that even the automatically annotated pairs follow this guideline: The symbolic systems always assume things as coreferent as possible.

Ambiguous Senses/Sentences. In this case, the controversy stems from lexical (polysemy) or syntactic ambiguity and depending on how we disambiguate, the inference label might be different. Example: *P: A person is folding a sheet. H: A person is folding a piece of paper*: Depending on whether *sheet* is taken to mean the bedsheet or a piece of paper, a different inference relation occurs. For such cases, we decide to disambiguate the ambiguous word with the sense that would possibly make the sentence coreferent to the other sentence (e.g., in this example, a piece of paper). Similarly, for an example with syntactic attachment ambiguity like *P: Two men are fighting in a cattle pen. H: Two men are fighting in a pen for cattle*, we assume the attachment that makes the two sentences coreferent.

Idioms with Prepositions. Idioms with prepositions can be problematic because a preposition might alter the meaning of the sentence. Example: *P: The couple is walking down the aisle. H: The couple is walking up the aisle*. This pair could be C if we take *walking up/down the aisle* to be antonyms. However, there is no clear definition of what is the *up/down* direction of an aisle and for some people the expressions are even synonymous and thus, the pair could also be an entailment. To this category also belong idioms such as *go up/down the path, burn up/down the house*, and so forth. For other similar idioms, for example, *go up/down the stairs/mountain*, it is clear that there is an upper and a lower side. So, we decide to label as contradictory pairs where there is a clear-cut distinction of the meaning and otherwise as neutral, assuming that the prepositions do not make a difference.

The categories presented so far already highlight some of the hard issues of the NLI task, but the challenges are solvable. In other words, it is clear that without an organized annotation effort the annotations will end up inconsistent, but if people agree on a specific treatment for each case, most problems are solved. However, clear guidelines

can only take us so far: The challenges posed by the following phenomena cannot be easily solved by specific guidelines in a purely single-labeled or certainty gradient task.

Phrasal Verbs. If a phrasal verb is used in the pair, the inference label could be different depending on what we take it to mean. Example: *P: A woman is picking a can. H: A woman is taking a can*. *Picking* prototypically means "selecting" and in this case, it does not necessarily entail physically taking the can. However, we could also more freely interpret *picking* as *picking up*, in which case there would be a clear entailment. No matter which approach is taken here, a single inference label cannot be representative enough of human inference.

Looseness. This concerns pairs for which it is not clear how strict or loose we should be with the meaning of specific words. Example: *P: A surfer is surfing a big wave. H: A surfer is surfing a huge wave*. *Big* is generally not the same as *huge* but in the context of a wave, they could probably be synonymous. Obviously, an NLI task with a single inference label or a certainty gradient is not optimal here because it forces us to choose between strict, logical inference and "people on the street" inference. Clearly there is no right or wrong here, nor certain vs. uncertain; there is merely a difference in the proportion of logic and common-sense the annotator chooses to use.

Gender/Age. A controversy arises about whether words like *man* and *woman* should be taken to refer to the gender or to the age of a person. Example: *P: A girl is eating a cupcake. H: A woman is not eating a cupcake*. A girl is a woman as far as the strict gender is concerned (at least in a binary gender approach), but not concerning her age. Again, this category makes clear how static the single inference label approach is: We need to decide between the strict "dictionary" definition, which would rather point to the gender, and the more common-sense approach, which would need to evaluate the context each time. Such a controversy does not concern human certainty/uncertainty, but rather a graded notion of logic and common-sense. As far as the gender is concerned, note that we also replace all occurrences of *lady* with the word *woman* as we find the term *lady* out-dated and almost sexist.[7]

Privative Adjectives. Such expressions occur when an adjective (or a noun) modifier contradicts the noun it modifies (Partee 2010), for example, *cartoon airplane, fake gun, animated bear*, and so on. Example: *P: An animated airplane is landing. H: The plane is landing*. Depending on whether the adjective is taken with its literal meaning or not, such inferences can be neutral or non-neutral. Once again, we are in a dilemma as far as these annotations are concerned: should we be strict judges that say that a cartoon airplane is not an airplane or go with the childish common-sense that allows cartoons to be real?

World Knowledge. Pairs requiring world knowledge to be solved also cause confusion and controversy. As with other categories, these pairs may have multiple right labels, depending on what we take the world to be like. Example: *P: A team is playing football. H: A team is playing soccer*. Logically speaking, *football* does not entail *soccer* because the US-based *football* is a completely different sport. But in Europe, the one does entail the other. A similar phenomenon occurs with a pair like *P: A man is cooking a breaded pork chop. H: A man is frying a breaded pork chop*. One doesn't need to be frying a pork chop, maybe

7 Since there is no corresponding *man* and *gentleman*.

they are just boiling it, but since we are talking about a *breaded* pork chop, chances are that there is frying taking place. This once again clearly shows how common-sense and logic are at play and how different annotations do not have to be opposing, but rather expressing different ends of the same spectrum.

3.2.2 Challenges. The issues listed in our guidelines already show the challenges of the current NLI task formulation. Very often, there is no single correct inference label. The label depends on what assumptions we make, particularly on whether we decide to judge things by their strict "dictionary" meaning or whether we allow for more everyday uses of words and meanings. This shows that the main reason the task is so error-prone is not the annotators not paying enough attention or being indecisive in their annotations, but rather annotators being forced to a decision expressing only one of their cognitive states: their logic or their common-sense. Thus, different people decide for different notions of logic/common-sense. This leads to noisy, inconsistent, and ill-formed datasets. Downstream, we find "confused" models that can neither generalize specific logic rules nor learn consistent common-sense patterns since the data they are training on contains both kinds of annotations. Evidence for this confusion is also given in Section 5.

Interestingly, this forced logic vs. common-sense decision of the annotators is independent of whether we decide to treat inference labels as discrete or gradable. Even if we treat the labels as gradable, this is a gradient of certainty: How probable it is that a pair is an entailment rather than a contradiction. But this probability is not independent of whether the pair is considered strictly logically or based on common-sense. Depending on this notion, the pair might in fact be an entailment *and* a contradiction at the same time. This can be seen in the following example: *P: The man is painting. H: The man is drawing*. If we are asked to judge how likely H would be, given P, we would probably say that there is 50% chance that this is an entailment, namely, that the man painting also entails drawing. But how do we arrive at this 50%? And why isn't it 40% or 60%? And why should the chance be 50% entailment and not 50% contradiction? And what is the remaining 50%? There is no satisfactory answer to these questions. However, if we are told to judge this pair according to our strict logic and our common-sense, we would be able to assign both inference labels, without contradicting ourselves: Logically speaking, the pair is a contradiction because at the very same moment, the man cannot be drawing and painting the same picture; according to our common sense, though, this could also be an entailment because (a) many people use the two terms interchangeably and (b) drawing usually precedes or complements painting. In fact, this could explain why the SICK corpus contains so many "asymmetrical" contradictions (Kalouli, Real, and de Paiva 2017a), that is, pairs that are not contradictory in both their directions A to B and B to A, as contradictions in logic should be: People annotating with different notions of logic and common-sense in mind. Thus, our proposed ordered sense space annotation is able to tackle these challenges and provide a solution that allows both cognitive states of logic and common-sense to be simultaneously expressed without falling into a contradiction.

4. The Ordered Sense Space Annotation

This section describes the proposed ordered sense space and details how the 15 annotation issues discussed can be treated based on this space. The section also provides a detailed presentation of the final re-annotated corpus.

4.1 Description of the Space

To solve the controversial inference issues presented above we propose an **ordered sense space** annotation. The space represents the different stances with which an annotator could label a given pair. On the one end of the ordered space, we find non-sensical inferences, namely, entailments, contradictions, or neutral relations that could never hold for a given pair; invalid inferences. This novel notion can be made clear with an example such as *P: The woman is walking. H: The woman is sitting*. Here, an entailing inference could never be valid and make sense because we could never infer that a woman is sitting, if we know that the woman is actually walking (assuming we mean the same woman and the same moment of time). On the other end of the ordered space, we have the notion of logical inference. This notion is to be understood as the strict, "textbook" kind of inference that a judge, a mathematician, or a logician would use to solve a legal/mathematical/logical problem. Thus, the two ends of this ordered space are rigid, concretely labeled points for which there should be complete human agreement. Between these two solid points, we find the notion of common-sense inference. This is the kind of inference that humans would be inclined to make in their everyday lives, without adhering to specific guidelines or restrictions, merely based on their knowledge about the world. This is a subjective and gradable notion, for which no human agreement can be pursued, and this is faithfully captured in the middle of the ordered space as a continuum; a continuum of common-sense, capturing potentially different common-sense labels of different people, inclining either toward the non-sensical or the logical end. The space is ordered because there are inferences that could be justified with common-sense, but then they would be almost non-sensical (thus positioned at the far left side of the common-sense continuum), and there are others that almost overlap in their common-sense and logical interpretation (thus positioned at the far right of the common-sense continuum); based on this judgment, the annotator can decide where exactly to put the label within the common-sense continuum.

With this, our proposed approach combines the traditional single-label method (captured at the ends of the ordered space) with the recently proposed graded approach (captured with the continuum in-between the ends of the space), and factors in the notion of *human sense* rather than uncertainty. The annotation process is as follows: Each pair is first represented by an unlabeled ordered sense space. Then, an annotator picks a **stance**, a point on the ordered space where they want to be. This stance might well depend on the pair that is to be annotated. For example, if the pair reads like a legalistic text, they might prefer to be on the logic end. If the pair is in colloquial language or talks about everyday items in a loose fashion, they might prefer to use common-sense. There is no privileged position. Alternatively, annotation guidelines could suggest a stance. Based on this stance, the annotator has to determine the label that fits best. They might try out each of labels E, C, and N to see which feels the best. The label they decide on should then be placed on the area of the space that represents the stance with which they made this decision. With this, each inference pair is represented by a labeled ordered sense space.

Note that an annotator might wish to assume both the common-sense and logic stances and think about things from both angles. This is indeed absolutely necessary for cases where no single inference label can be determined, as in the examples we discussed in Section 3.2.1 and that can now be solved with this annotation scheme. This feature is also our main contribution to the matter of annotation as we permitted annotators (ourselves) to use labels that explicitly express a certain cognitive stance.

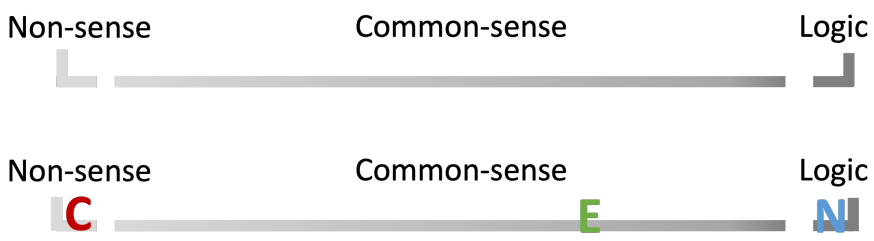

Figure 3
The proposed ordered sense space annotation. Each pair is represented by the ordered space, on which each of the inference relations (entailment, contradiction, neutrality) may be placed. The annotator has to decide on the stance they wish to take and decide for the label (E, C, N) that fits best to it. The ordered space above shows the annotation for the pair *P: A group of people is on a beach. H: A group of people is near the sea.*

On the other hand, annotators would not want to assume the non-sensical stance during their annotation, at least as far as we can tell—maybe only in tasks where it is required to explicitly state that a specific inference is outside the realm of logic and common-sense.

Figure 3 illustrates the proposed approach for the pair *P: A group of people is on a beach. H: A group of people is near the sea.* The traditional single-label approach will struggle with this pair because the pair will end up annotated as neutral by some annotators and as entailment by others. According to Pavlick and Kwiatkowski (2019), this ambiguity effect is not annotation noise and does not disappear even if we ask many more annotators, because it stems from the inherent human disagreement on whether beaches are "prototypically" meant to be by the sea (and not a lake, a river, etc.). On the other hand, a graded or ordinal approach, as the ones proposed by Chen et al. (2020) and Zhang et al. (2017), respectively, would not solve the uncertainty either: The probability of entailment or neutral would be somehow distributed, without further information on what drives this distribution and what it expresses. Our proposed approach is able to combine these two methods and make the best of them. First, we decide whether to be strictly logical reasoners or whether to estimate what the common-sense individual would say. In the logical sense, a beach does not always indicate the sea: Lakes and rivers occasionally have beaches. From this stance, the label would be N. From the common-sense side, it seems to be an entailment E. However, there is no strict standard for the common-sense label the way there is for a logical one. An annotator would either give their own "gut feeling" or try to estimate what the gut feeling of others would be. At this time, we have not explored whether there would be a difference in results based on these two options. For completeness and better clarity, in Figure 3 we also show how the contradiction label would be at the non-sensical end, should one want to take this stance: Being at the beach can never contradict being near the sea. The non-sensical stance will not be assumed for further illustrations or annotations, unless otherwise described.

Note that in the proposed approach the same inference relation can be found in more than one area of the ordered space, for example, when the logical interpretation coincides with the common-sense one. Consider the pair *P: Three boys are jumping in the leaves. H: Three kids are jumping in the leaves*, represented by the ordered sense space illustrated in Figure 4. Here, the entailment relation is valid both based on a strict,

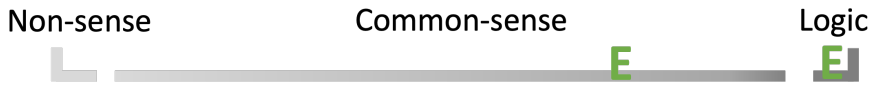

Figure 4
The proposed annotation of the pair *P: Three boys are jumping in the leaves. H: Three kids are jumping in the leaves*. Here, the entailment relation is located both on the logical end and within the common-sense area.

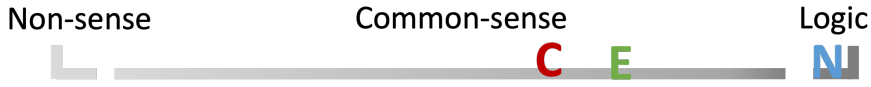

Figure 5
The proposed annotation of the pair *P: A person is kicking a ball for a soccer game between somebody's feet. H: A person is kicking a soccer ball between their feet*. Here, entailment and contradiction are both located within the common-sense continuum, accounting for the different common-sense notions of different people.

logical interpretation and on common-sense reasoning. Thus, the space features the E label twice.

Interestingly, our proposed approach allows for even more complex annotation cases. Consider the pair *P: A person is kicking a ball for a soccer game between somebody's feet. H: A person is kicking a soccer ball between their feet*. Assuming a logical stance, we would assign the neutral label because *somebody* is ambiguous in that we do not know whether their own feet or somebody else's feet is meant. But from a common-sense point-of-view, the decision is not as easy: Depending on the different disambiguations of *somebody*, both the entailment and the contradiction label may apply (entailment if *somebody* and *a person* corefer, and contradiction if they do not). The annotation is shown in Figure 5. But this setting can be efficiently served by the continuum of our ordered sense space, which captures the distribution of human interpretations based on common-sense and thus allows for different and even opposing annotations within the common-sense continuum.

Another category of complex cases that is treated by our approach contains examples whose common-sense label is very close to the non-sensical end of the space. An example is the pair *P: A monkey is riding a bike. H: A bike is being ridden over a monkey*, illustrated in Figure 6. Clearly, if thinking strictly logically, this pair should have a neutral relation because from P we cannot say anything about H. Most people would probably also say that entailment and contradiction are here entirely non-sensical. But apparently not everybody: Some of our annotators thought that, taking a common-sense stance, contradiction might be the right label, because if we are talking about the same monkey, it can be either on the bike or under the bike but not both at the same time (thus, contradiction). Still, these annotators agreed that this is a rather far-fetched case of contradiction (because how probable is it that anybody is riding in the air over a monkey?) and thus placed the contradiction label at the left end of the common-sense continuum, just before the non-sensical end. Thus, such examples show

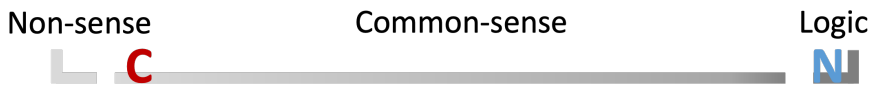

Figure 6
The proposed annotation of the pair *P: A monkey is riding a bike. H: A bike is being ridden over a monkey*. From a strictly logical point of view, this pair is neutral. However, for some annotators the contradiction label might still be within the realm of common-sense, though almost non-sensical.

how the proposed sense space, being ordered, can even deal with such corner cases. Additionally, such pairs justify the need for the novel notion of non-sensical inferences.

By being novel in its formulation, the proposed annotation scheme might at first sight seem too complex for crowdsourced annotations. However, the proposed annotation is in fact much simpler than detailed annotation guidelines of other works, and more reliable than efforts without any guidelines. Essentially, the annotators are only asked to judge the pairs based on their intuitive logic or common-sense. They do not need to follow long annotation guidelines nor to perform a non-natural task. At the same time, the nature of the proposed scheme integrating the specific stance of the annotator avoids well-known problems of annotation efforts, which do not provide any specific guidelines, for example, SNLI (cf. Marelli et al. 2014b; Kalouli, Real, and de Paiva 2017b; Kalouli et al. 2019, 2021). Thus, the proposed annotation scheme can easily scale-up to crowdsourcing annotators. Despite the nature of the proposed space, which allows for the human inherent inference variability, concerns about the difficulty of detecting low-quality workers and measuring the inter-annotator agreement (IAA) are not justified. Low-quality workers can still be detected and IAA can be measured by looking at the ends of the space, which represent solid points with the same labels for almost all annotators, as detailed above (see also Section 4.3, where we report on IAA). We conduct a crowdsourcing experiment in Section 6.

4.2 How the Proposed Annotation Cures the Maladies of NLI Datasets

Based on the proposed ordered sense space, the 15 annotation issues discussed can be revised to solve the open challenges and to also offer better solutions to some of the compromises we needed to make before (see Section 3.2.1). These updated guidelines can also be found in Appendix A. Note that our manual annotation based on the proposed sense space is not supposed to reflect the actual distribution of crowdsourced annotations for the SICK corpus (like in Pavlick and Kwiatkowski 2019); rather, it aims at a more reliable annotation of the corpus, allowing for and dealing with the inherent human variability. Particularly, our manual annotation based on the proposed space should serve as a proof for the usefulness and efficiency of the proposal, which solves the aforementioned issues. In fact, we are confident that the flexibility and modularity of the proposed scheme is such that further issues that do not fall into the aforementioned categories can be efficiently addressed too.

For the categories of ungrammatical and semantically absurd pairs and for the aspect/tense phenomenon, no new treatment is initiated because these issues do not concern the annotation per se.

For the coreference category, our proposed approach allows for more flexibility. Recall that in our initial guidelines we were forced to give one single label to pairs for which the coreference was not clear, for example, assuming that, if the two events are highly unlikely to happen at the same time, then they should be contradictory; now, we can allow for two distinct labels based on the stance we choose to take. In the example *P: A man is dancing. H: A man is speaking*, we can assume that, if we judge the inference according to the strict guidelines we defined before, a contradictory relation is justified. At the same time, the natural label we would be assigning with our common-sense is the neutral label (many people talk while dancing—e.g., to their dance partner).

For ambiguous senses and sentences we are also not forced to make a compromise. For the example *P: A person is folding a sheet. H: A person is folding a piece of paper*, we can assign up to three labels: the pair is neutral if we stick to logical inference (we just do not know if we are talking about a piece of paper or bed linen), but for the common-sense it could be an entailment, if we decide to take *sheet* to be a piece of paper, or even a contradiction if we take *sheet* to mean *bed linen* and thus, form a contradictory relation.

Idioms with prepositions can now also get a better treatment. To a pair such as *P: A couple who have just got married are walking down the aisle. H: The bride and the groom are leaving after the wedding*, we can assign two labels: Strictly speaking, this inference is neutral because we cannot say whether *walking down* actually means walking toward the door of the ceremony location and thus leaving; it could also be the case that the couple walks down the front part of the ceremony location to pick up some of the papers. However, in common-sense terms, this inference is probably an entailment because we know that the couple is already married and thus, chances are that they are actually leaving the place. Such examples confirm the flexibility and consistency that the proposed sense annotation allows for. In fact, for the remaining categories the sense annotation does not only offer higher flexibility, but actual solutions.

For the phrasal verbs, we do not have to pick one single label and thus we can be sure that our annotation is representative of human inference. For a pair like *P: A man is cutting a tree with an axe. H: A person is chopping down a tree with an axe*, we are at the liberty of assigning two labels to express the whole spectrum of possible human inferences: strictly speaking *cutting* does not entail *chopping down*, even when speaking of an axe and a tree, so this should be neutral. Nevertheless, for most people's commonsense, the phrasal verb can be used interchangeably with the other verb and thus, an entailment relation can be established.

For the challenging category of looseness, the proposed annotation is a promising way out. Consider the example *P: A man is pushing the buttons of a microwave. H: One man is powering the microwave.* Strictly speaking, this is a neutral relation: One might be pushing buttons without actually turning on the appliance; maybe they are just setting the time. However, most people would be happy to accept this relation as an entailment because in most scenarios the pushing of buttons has the goal of powering the appliance. The same logic applies for the pair *P: A woman is squeezing a lemon. H: The woman is squeezing juice out of a lemon.* Now, squeezing a lemon does not necessarily mean that one is actually squeezing the juice out of it; maybe somebody really just presses the lemon to see how ripe it is. Thus, the neutral label is located at the logic end of the space. However, in most daily scenarios, one refers to squeezing a lemon to mean getting the juice out of it, and thus the entailment label should be placed in the common-sense range of the continuum (more to the right, according to our own intuition, which is, however, subjective and refinable at all times).

Moving on to the gender/age challenge, we can examine the example *P: A cupcake is being eaten by a girl. H: A woman is eating a cupcake*. If we decide to use the dictionary

definitions of things, the entailment relation must be placed at the logic end of the space. Taking genders into account, a girl is a woman. But most people would not be happy with this annotation. This is because the context of a girl eating a cupcake probably brings up the picture of a child and less of an adult. Thus, our common-sense would be inclined to judge this relation as neutral.

Privative constructions can also be better handled with the proposed ordered sense space because it allows for the context-sensitivity required. Consider the example we saw before: *P: The cartoon airplane is landing. H: The plane is landing.* If a cartoon airplane is landing, there is certainly no real airplane landing. Thus, strictly speaking the relation among the pairs is a neutral one. But what if we situate the pair in a more casual context, maybe involving children, where we are discussing that cartoon airplanes can land, take off, and so on? Then, the pair should clearly be entailing. The sense space offers this duality.

Last, we discussed the challenges posed by pairs requiring world-knowledge. This is one of the most significant categories as state-of-the-art models are claimed to be able to capture it well. Thus, it becomes important to have reliable annotations, representative of the phenomenon. This can be seen in the example mentioned before: *P: A man is cooking a breaded pork chop. H: A man is frying a breaded pork chop*. The neutral label can be placed at the logic end of the space: One does not necessarily need to be frying, if they are cooking. However, most people will say that if something is *breaded*, it is most probably fried. Thus, they would be happy to place the entailment label within the common-sense continuum, probably even toward the right, logical end of the space.

4.3 Resulting Corpus

Based on the proposed sense annotation scheme we are able to manually annotate all 3,138 pairs that could not be reliably annotated by the automatic systems. Specifically, each of the 3,138 pairs is annotated by 3 from our group independently, and each annotator is free to choose one or both stances for each pair. For the logical end of the ordered space, we take the majority vote of the annotations because we want to end up with one common, uniform label—recall that this notion is supposed to be a rigid, single-labeled point in space. For the logical category it is also useful to compute the IAA, which lies at 0.635 Fleiss kappa for the manually annotated pairs and at 0.835 for all pairs (substantial and almost perfect agreement, respectively). In contrast, for the common-sense continuum we do not require such uniformity because of the very nature of the continuum. It is also meaningless to try to calculate the IAA since we explicitly allow and expect disagreement.

Out of the 3,138 pairs, 307 have different labels for each of the two stances, that is, around 10% have different labels in the logical end and the common-sense continuum. This percentage might give the initial impression that the majority of pairs can be fully captured by a single label and that, thus, the more complex annotation scheme is not necessary. However, as detailed in Section 5, this relatively low percentage makes a significant difference for the model training and testing. Table 1 presents the distribution of the labels of the manually annotated pairs, also in comparison to the original SICK labels. Note that the table presents the common-sense labels as distinct labels (rather than different labels on a continuum) and this might seem to go against our own argumentation. The reason for this is that from the 3,138 pairs that were manually annotated, only 15 of them had different common-sense labels; all others might have had a distinct logic and common-sense label, but the common-sense label itself was consistent among the annotators. Thus, we refrained from using this small percentage of pairs for further

Table 1
Distribution of the manually annotated pairs based on the proposed ordered sense space annotation as per original inference label. AB stands for the semantically absurd pairs and UG for the ungrammatical ones.

		Logic Stance					Common-sense Stance				
		E	C	N	AB	UG	E	C	N	AB	UG
	E	1,009	36	241	14	29	1,128	37	121	14	29
SICK original	C	5	426	34	1	12	6	433	26	1	12
	N	49	699	527	26	30	84	713	478	26	30
Total per new label		1,063	1,161	802	41	71	1,218	1,183	625	41	71
Total	**3,138**	**3,138**					**3,138**				

experimentation as it would not be able to offer any conclusive evidence. However, this should not suggest that the representation of common-sense as a continuum is not necessary: if more pairs are annotated and especially if the task is performed by people of very different backgrounds, for example, crowdsourcing workers, the common-sense continuum can exactly capture the expected variation.

We make several observations from this table. First, the entailments and the contradictions of the original corpus agree more with the label of the common-sense stance than of the logical stance. From the entailments of the original corpus, there are 1,128 pairs, which we also annotate as E based on common-sense, vs. 1,009 which we annotate as E based on logic. Similarly, from the contradictions of the original corpus, there are 433 pairs which we also annotate as C based on common-sense vs. 426 based on logic. This is not surprising as most SICK annotators used common-sense to judge the pairs and did not think of more complex, logical rules. However, the original neutrals show the opposite picture: They agree more with the logic (527) than with the common-sense (478) annotation. This is consistent with the logic of our sense annotation. Based on strict logic, many pairs get a neutral relation, but, judged more leniently, based on common-sense, several of these pairs get an entailment or a contradiction label. Thus, the pairs judged as neutral based on common-sense are a proper subset of the pairs judged as neutral based on logic. Additionally, the table shows that for entailments and contradictions the new annotation mostly agrees with the original annotation, both for the logical and the common-sense viewpoint. In contrast, for neutrals there is more disagreement: Based on the logic stance, 699 pairs are annotated as C, instead of N, and based on common-sense, 713. This shows that 12.4% and 12.7%, respectively, of the original neutral pairs are annotated in a different way. Another interesting observation is that in the logical annotation there are more neutrals than in the common-sense annotation (802 vs. 625), but fewer entailments and contradictions (1,063 vs. 1,218 and 1,161 vs. 1,183, respectively). This reflects the notion of our proposed annotation: When strict logic is required, more pairs get to be neutral than when common-sense is used. Conversely, when common-sense is used, more pairs get to be entailments and contradictions than when logic is used. More generally, out of the 3,138 pairs we annotated manually, 1,176 (37.4%) have a different label than the original label within the logic stance and 1,099 (35%) within the common-sense stance.[8]

8 Assuming that the common-sense labels are merged based on the majority vote; see explanation above.

Table 2
Distribution of the entire corpus based on the proposed ordered sense space annotation as per original inference label. AB stands for the semantically absurd pairs and UG for the ungrammatical ones.

		Logic Stance					Common-sense Stance				
		E	C	N	AB	UG	E	C	N	AB	UG
	E	2,537	36	241	14	29	2,656	37	121	14	29
SICK original	C	5	1,413	34	1	12	6	1,420	26	1	12
	N	49	699	4,801	26	30	84	713	4,752	26	30
Total per new label		2,591	2,148	5,076	41	71	2,753	2,173	4,900	41	71
Total	**9,927**	**9,927**					**9,927**				

These percentages are different if they are calculated for the entire corpus. Because two-thirds of the corpus' pairs are annotated automatically with the help of the symbolic systems, most of them only have a single-label annotation. However, this single-label annotation can easily be transformed to our proposed ordered space annotation. Recall that the automatic process only considers reliable the labels on which the symbolic systems *and* the original SICK annotation agree. Thus, these labels reflect well both the logic and the common-sense interpretation: The logic interpretation is captured by the symbolic systems and the common-sense by the SICK original annotators. So, the pairs that are automatically annotated and only have a single label can be taken to have this label both from the logic and the common-sense stance. With this, the distribution of labels over the entire corpus develops as shown in Table 2. For this table we can observe the same tendencies as above. Overall, around 9% of the entire corpus has *different* annotations in the original corpus and our re-annotated version (both with respect to the logic stance and the common-sense one). This percentage is lower than the one for the manually annotated pairs because this distribution contains more of the automatically annotated pairs which by definition agree with the original label.

At this point, it is also interesting to see whether and how the new annotation labels correlate with specific phenomena of our taxonomy categories. To this end, we provide the taxonomic category and the annotation labels for the 345 pairs that we annotated for calibration purposes.[9] Table 3 summarizes the results. The column marked with # represents the number of pairs found for each category, while the column *distinct* expresses what *percentage* of these pairs was annotated with two distinct labels for logic and common-sense. The rest of the columns give the percentage of pairs that was annotated with a specific (distinct) label combination for logic and common-sense, respectively. For example, the column *N/E* tells us that all pairs (100%) that were marked with aspect/tense issues and were given two labels got the annotation N for the logic stance and the annotation E for the common-sense stance. Column *N/C,E* gives the percentage of the pairs that were given two distinct labels for logic and common-sense and for the common-sense stance there was more than one label (i.e., C and E). The table shows us that the taxonomic issues correlate heavily with the pairs being given distinct labels for the logic and the common-sense stances. All categories of our taxonomy with at least one occurrence within the calibration set were per majority annotated with two distinct labels for logic and common-sense, for example, almost

9 The taxonomic category is only available for these pairs and not for the whole corpus.

Table 3
Calibration pairs (345): Correlation of the taxonomic category with two distinct labels for logic and common-sense and with a specific label combination. All columns except for # represent percentages, e.g., the *distinct* column gives the percentage of pairs that have two distinct labels for logic and common-sense, while a column such as *N/C* represents the percentage of pairs that have N as their logic label and E as their common-sense label. Column # represents the absolute number of pairs found in each category. Columns with two labels after the / (e.g., *N/C,E*) represent cases where more than one label was given for common-sense (according to the proposed continuum).

Category	#	distinct	N/C	N/E	N/C,E	C/E	E/C	C/N	E/N	C/E,N
ungrammatical	3	0	–	–	–	–	–	–	–	–
non-sensical	1	0	–	–	–	–	–	–	–	–
aspect/tense	5	60	–	100	–	–	–	–	–	–
coreference	4	75	–	66.6	–	–	–	33.3	–	–
ambiguous	31	96.7	3.3	73.3	10	13.3	3.3	–	–	–
idioms	0	0	–	–	–	–	–	–	–	–
phrasal verbs	4	50	50	50	–	–	–	–	–	–
looseness	20	90	–	61.1	16.6	22.2	–	–	–	–
gender/age	12	91.6	–	9	–	9	45.4	–	18.1	18.1
privative	0	0	–	–	–	–	–	–	–	–
world knowledge	11	90	–	80	–	20	–	–	–	–
"normal" (all others)	254	0	–	–	–	–	–	–	–	–

97% of the ambiguous cases and 92% of the gender/age pairs were given two distinct labels. On the contrary, "normal" pairs, that is, pairs for which no special taxonomic category could be established, were never given two distinct labels. This suggests that for the majority of the pairs, the logic and the common-sense stance coincide and that, thus, existing annotation schemes can efficiently capture them, giving the impression of being accurate. However, we see that all other "harder" cases can hardly be captured with a single label or a graded scale annotation. In particular, we observe that such pairs mostly get an N/E annotation, that is, they are dimmed neutral according to logic and entailment according to common-sense. This nicely captures the intuition behind the proposed scheme: If strictly judged with logic, many pairs have to be neutral; however, if the more lax common-sense is allowed, most people will be happy to assume that these pairs are entailments.

5. Experiments

Although the proposed ordered sense space annotation is theoretically able to handle well the challenges of the NLI task, as explained above, it is also important to investigate how state-of-the-art neural models behave practically with respect to the ordered sense space annotation. In particular, we want to address the following research questions:

- How do models perform on the newly annotated corpus when the logic and the common-sense annotation are used as the gold-standard and also in comparison to their performance on the original corpus?
- Does fine-tuning on the new labels of the logic and common-sense stance alter the performance of the models on the original and the re-annotated corpus?

- How do models perform when tested only on the pairs where the label based on logic and on common-sense differ from the original SICK label, respectively? Is there a significant performance difference that can tell us more about the nature of these pairs? And how well can symbolic systems deal with such pairs?
- What is the performance of the models when tested only on the pairs that have different labels for the logic and common-sense stance? Are such pairs with inherent variability more difficult for models?
- Do the probabilities predicted by the models reflect the different sense stances for pairs that have different labels for the logic and common-sense stance? In other words, are models able to predict the annotation space as proposed within this work?
- How do models trained on the original and the new version of SICK behave when confronted with entirely different test data, for example, from corpora such as SNLI and MNLI?
- The ordered sense space was proposed for humans. Do the findings here give us reason to think that it is a useful way to think about models and what they are doing?

To answer these questions we perform several experiments in different settings, namely, with different training and test sets. All experiments are performed with the BERT (base) (Devlin et al. 2019) and the RoBERTa (Liu et al. 2019b) pretrained models in the HuggingFace implementation (Wolf et al. 2020). BERT is chosen as a representative model of the transformer architecture and because it has shown a considerable performance jump compared with previous neural models in many different NLI datasets. RoBERTa is chosen because it has proven a very strong model for NLI, outperforming BERT. For all experiments we fine-tune the parameters of learning rate and epochs based on the setting that delivers the best performance. The best performance is calculated based on the average validation loss across 5 runs with different seeds.

5.1 Experimental Set-ups

The experiments include three training (fine-tuning) settings and 12 testing sets. The necessity for these different experimental settings becomes clear in Section 5.2 during the discussion of the results. The SICK corpus is split into its standard train/validation/test splits, as these were defined during the SemEval 2014 challenge (Marelli et al. 2014a). The train set contains 4,500 pairs, the validation set 500 pairs, and the test set 4,927 pairs. For training, the pretrained models are fine-tuned on the training set (a) annotated with the original labels, (b) annotated with the labels of the logic stance, and (c) annotated with the labels of the common-sense stance. For testing the different models, the following test sets are considered. First, we test on the SICK test set annotated with the original labels. Next, we test on the test set annotated with the labels of the logic stance and on the test set annotated with the labels of the common-sense stance.[10] Additionally, we test on that subset of the test set for which the original label agrees with our new label.

10 See note in Section 4.3 on which labels are exactly used for the common-sense stance.

Two settings are considered. In the first one, the subset contains pairs whose original annotation agrees with the label of logic (4,352 pairs), while in the second setting the subset contains pairs whose original annotation agrees with the label of common-sense (4,388 pairs). Furthermore, we test on the complements of these subsets. More precisely, we test on the subset of the test set that contains pairs whose original annotation does *not* agree with the label of logic (575 pairs) and on the subset with pairs whose original annotation does *not* agree with the label of common-sense (539 pairs). A further experimental setting considers the subset of the test set that contains pairs that have distinct labels based on the logic and common-sense stances. This subset contains 149 pairs and again two settings are distinguished: in the first one the pairs are annotated with the label of logic, while in the second one they are annotated with the label of common-sense. In this setting, we also test whether the models are able to predict the whole spectrum of the proposed annotation space, by predicting the logic and the common-sense labels of these pairs. Finally, we also test on the SNLI (Bowman et al. 2015) test and MNLI (Williams, Nangia, and Bowman 2018) development set.

5.2 Results

Testing on SICK Test Set. The experiments and the results of testing on the SICK test set are shown in Table 4. The table shows the accuracy of the BERT and RoBERTa models when fine-tuned and tested on the different settings.

First, the table confirms previous literature findings on these two models. We see that BERT and RoBERTa are able to achieve an accuracy of 85.3% and 90.5%, respectively, when trained on the original SICK training set and tested on the original SICK test set. This performance is comparable to the performance reported by Devlin et al. (2019) and Liu et al. (2019b) for SNLI and MNLI and shows that the smaller size of SICK does not have a negative impact on performance. This also confirms that our findings are transferable to larger NLI corpora. However, when the models are tested on the SICK test set annotated with the logical or the common-sense stance, performance drops to around 80%–83%, respectively. This is not surprising since the original SICK training set is not suitable for capturing the logic and common-sense nuances.

When training on a dataset with a specific stance, BERT and RoBERTa show different behaviors. RoBERTa performs best when trained and tested on datasets of the same stance, as expected. For example, when trained on the SICK train set annotated with the common-sense stance, RoBERTa achieves 87.4% on the test set annotated with the same stance, while only 86.9% and 85.4% for the logic stance and the original dataset,

Table 4
Accuracy of the pretrained models when fine-tuned and tested on different train and test sets. The notation *.logic/.common-sense/.original* states which stance the annotator assumed when providing the label.

Testset	**Trainset**					
	original SICK.train		SICK.train.logic		SICK.train.common-sense	
	BERT	**RoBERTA**	**BERT**	**RoBERTA**	**BERT**	**RoBERTA**
original SICK.test	85.3	**90.5**	82.9	85.6	82	85.4
SICK.test.logic	80.6	83.9	81.4	**87.9**	80.3	86.9
SICK.test.common-sense	80.8	84.2	81.2	87.6	80.7	**87.4**

respectively. This clearly suggests that RoBERTa is able to capture much of the notion of logic/common-sense that is encoded in these differently annotated datasets. On the contrary, BERT shows a surprising behavior at first sight. Even when trained on the SICK training set annotated with the logic or the common-sense stances, BERT still performs better on the original SICK test set than on the newly annotated SICK test sets, respectively. This would mean that BERT does not learn to efficiently distinguish the specific notions of logical and common-sense inference only based on SICK. Particularly, this could suggest that the specific BERT architecture does not allow the model to improve only based on a minority of instances. Recall that after annotating the whole corpus, we found that 9% of it was differently annotated by us and by the original annotators. Thus, because most of the new annotations coincide with the original ones, there are only a few different ones and these do not seem to get any special treatment by BERT so that their effect is "absorbed" by the rest of the pairs, that is, they stand in the minority and thus cannot contribute to a high learning attention. Thus, BERT still performs best on the original SICK test set.

Testing on the "Agreeing/Disagreeing" Subsets of the SICK Test Set. To confirm these findings and to also investigate further the "absorbing" effect of the different annotations, we perform experiments on subsets of the SICK test set. Specifically, we split the SICK test set to pairs that have the same original label as our label and pairs that have a different original label from our label. In each of these splits, we consider two settings. The first considers pairs whose original label is the same/different from our label based on logic, while the second setting includes pairs whose original label is the same/different from our label based on common-sense. An example of each category is given in (1)–(4).

(1) **Original label (C) is the same as the label based on logic (C):**
P: Two dogs are wrestling and hugging.
H: There is no dog wrestling and hugging.

(2) **Original label (N) is different from the label based on logic (C):**
P: A man, a woman and two girls are walking on the beach.
H: A man, a woman and two girls are sitting on the beach.

(3) **Original label (E) is the same as the label based on common-sense (E):**
P: An onion is being cut by a man.
H: A man is slicing an onion.

(4) **Original label (C) is different from the label based on common-sense (N):**
P: A man is laughing.
H: A man is crying.

Example (1) is judged as a contradiction by the original annotators. In terms of a strict, logical inference, this pair is indeed a C and is thus annotated by us the same way. (2) is an opposite example: Here, the annotators considered the pair neutral, probably thinking that the group of people could be taking some breaks while walking on the beach. However, according to our definition of strict, logical inference, this pair is C because at the very same moment, the very same people cannot be walking and sitting. Moving on to the common-sense labels, the original annotators and us agree on example (3). The pair is labeled an E because within common-sense, cutting is (almost) synonymous to slicing and especially for vegetables, cutting pretty much entails slicing of some kind. For example (4), our label does not agree with the original label. Strictly speaking, the pair is neutral because a person who is laughing might indeed be having tears at the

Table 5
Accuracy of the the pretrained models when fine-tuned and tested on the "agreeing/disagreeing" subsets of the SICK test set. The notation *.logic/.common-sense/.original* states which stance the annotator assumed when providing the label.

Testset	Trainset					
	original SICK.train		SICK.train.logic		SICK.train.common-sense	
	BERT	**RoBERTA**	**BERT**	**RoBERTA**	**BERT**	**RoBERTA**
pairs where SICK.test.logic == SICK.test.original	88.7	92.4	88.3	92	87.5	91.5
pairs where SICK.test.common-sense == SICK.test.original	88.6	92.3	88.1	91.5	87.5	91.4
pairs where SICK.test.logic != SICK.test.original	20.1	19.5	29.4	57.2	26.1	52.7
pairs where SICK.test.common-sense != SICK.test.original	18.8	18.7	26.5	55.4	25.8	54.5

same time, either because the joke was so good or because they are so touched. This is probably also what the original annotators thought. However, in a more everyday scenario, somebody who is laughing is most often not crying and thus, the pair could be a contradiction within common-sense.

The results of these experiments are shown in Table 5. The first major observation is that the performance of the models on the pairs where the original label is different from the new label is much lower than on the pairs with the same label. For the former, the performance is at most at high 20s for BERT and 50s for RoBERTA, while for the latter the performance is at high 80s and 90s, respectively. This tremendous performance difference allows for a couple of conclusions. First, it confirms the finding that the common practice of using standard (train/dev/test) splits is not necessarily representative of the models' performance because testing with different splits—as the ones we choose here—does not allow for the reproducibility of the (high) performance (see Gorman and Bedrick 2019; van der Goot 2021). In particular, it shows that low performance might be so well "hidden" by the high performance of the standard test split that it goes unnoticed. Second, this large performance gap indicates that the pairs that have been annotated differently by the original annotators and by us are inherently difficult and ambiguous to annotate, and thus trained models cannot deal with them as effectively. Interestingly, this difficulty effect does not disappear even if we consider as the gold label the label of the common-sense stance, which is supposed to capture this inherent variability. This finding confirms previous literature (Pavlick and Kwiatkowski 2019; Chen et al. 2020) and highlights the necessity of the proposed ordered sense space because the traditional single-label task cannot even capture this inherent variability.

The difficulty of these pairs can also be seen when running them on the symbolic systems. As shown in Table 6, the symbolic systems have an accuracy of 78% to 83% for the pairs on which the original label is the same as our label based on logic, but can reach at most 40% when evaluated on the pairs where the labels are different. This confirms that some pairs are not only hard for neural models like BERT and RoBERTa, but even for logic-based systems. Here, a further interesting observation is that the symbolic systems achieve a higher performance than BERT and RoBERTa for the pairs whose original label is different from the label based on logic (last row of Table 6 and the second to last row of Table 5). First, this is in line with our intuition about the nature of the symbolic systems being reliable on logical inferences. Additionally, it highlights

Table 6
Accuracy of the symbolic systems when tested on different test sets.

Testset	Symbolic Systems				
	ccg2lambda1	**ccg2lambda2**	**LangPro**	**MonaLog**	**GKR4NLI**
original SICK.test	80.5	81.1	80.5	75.3	77
SICK.test.logic	75.6	76.3	75.5	74.3	78.4
SICK.test.common-sense	74	74.4	74	72.5	77.5
pairs where SICK.test.logic == SICK.test.original	82.5	82.9	82.4	78.6	83.2
pairs where SICK.test.logic != SICK.test.original	23.3	25.9	23.4	41.7	42

the need for training data capturing the fine nuances of logical and common-sense inference, if we are aiming at human-like models.[11]

Going back to Table 5, we can see that the low performance of the models on the pairs with different original and new labels is not significantly affected by the training set. Training on the new re-annotated sets only slightly improves accuracy: Training on the SICK training set annotated with the logic stance does deliver a higher accuracy on the pairs whose original label is different from the label based on logic, when compared to training on the original SICK or the SICK train set annotated with the common-sense stance—for both models. Conversely, training on the SICK train set annotated with common-sense does not deliver better performance on the pairs whose original label is different from the label based on common-sense, than on the pairs whose original label is different from the label based on logic. This might seem counter-intuitive at first sight, but can be explained by the nature of the pairs. The pairs that have different original and common-sense labels are a subset of the pairs that have different original and logical labels because in many cases the two labels coincide, as discussed above.[12] Thus, training on the SICK train set annotated with common-sense shows comparable performance on the pairs whose original label is different from the label based on common-sense and on the pairs whose original label is different from the label based on logic. Overall, we also make the same observation as before: RoBERTa is better than BERT in dealing with pairs where the original and our label are different; RoBERTa can handle harder pairs more efficiently. This confirms the literature finding (Liu et al. 2019b) that RoBERTA outperforms BERT on tasks of inference.

Testing on the Subset of the SICK Test Set with "Distinct" Labels. The subset splitting we perform sheds light on pairs with inherent inference difficulty and on the corresponding model performance and suggests the need for the proposed ordered sense space annotation. To further investigate the effect of the sense space, we consider another group of "problematic" pairs: pairs that have two distinct labels, that is, based on logic they have a different annotation than based on common-sense. Examples of such pairs can be found in Section 4.2. Here, we create a subset of the test set containing only those

11 The precision and recall of the symbolic systems are given in Table 10 in Appendix B.
12 In other words, all pairs have a label based on logic, but not all pairs have a *distinct* label based on common-sense, because in many cases the two are the same.

Table 7
Accuracy of the pretrained models when fine-tuned and tested on the subset of the SICK test set with "distinct" labels . *.logic/.common-sense/.original* states which stance the annotator assumed when providing the label.

Testset	**Trainset**					
	original SICK.train		SICK.train.logic		SICK.train.common-sense	
	BERT	**RoBERTA**	**BERT**	**RoBERTA**	**BERT**	**RoBERTA**
pairs with distinct labels (logic annotation)	38.5	36.2	46.4	**49.8**	35.7	35.5
pairs with distinct labels (common-sense annotation)	46	48.4	37.5	38.1	48	**49.5**

pairs that have distinct labels for the logic and common-sense stance and, again, we experiment with two settings: In one setting the label based on logic is taken as the gold label and in the other, the label based on common-sense is considered gold.

Table 7 shows the results. When testing these pairs on their logic annotation, we see that training on the label based on logic of the SICK training set gives the best results, as expected. Training on the original SICK labels gives the next best results for both models. This is also expected because the original labels are somewhere between the logical and the common-sense stance: In some cases, annotators tried to be very precise and "logical" and in some others, common-sense prevailed. Overall, there seems to be a continuum of performance: performance on common-sense $<$ performance on SICK_original $<$ performance on logic. We get a similar but inverse picture when testing on the label based on common-sense of these pairs. The best performance is achieved when training on the common-sense annotation of the SICK training set, as expected. There is a slightly worse performance when training on the original BERT, and the worst performance is when training on the label based on logic. Again, this is according to our intuition: Training on the original SICK gives us better performance than training on the label based on logic, because the original label captures more of common-sense than the label based on logic does. Here, the continuum described previously is inverted: performance on logic $<$ performance on SICK_original $<$ performance on common-sense. Overall and across train and test sets, we can observe the same picture as for the subsets containing pairs with different original and new labels: The performance on the pairs with distinct labels is much lower than the performance on the entire test sets. This confirms the finding of the inherent difficulty of certain pairs.

Predicting All Elements of the Annotation Scheme for the SICK Test Set with "Distinct" Labels. Up until now, we have only evaluated settings in which the model prediction is compared either with the logic or the common-sense label. However, because our goal is the creation of a dataset with more realistic labels and the use of the proposed annotation scheme in NLI tasks, it is important to investigate whether the models can predict the whole spectrum of the annotation space. In other words, we need to investigate whether the models are able to learn the two different stances and reflect them in the predicted probabilities. For this purpose, in this evaluation setting we do not consider the predicted label of the model with the highest probability, but we look at the whole probability distribution. With this, we want to see whether this distribution reflects the different stances captured in the proposed annotation space, that is, whether

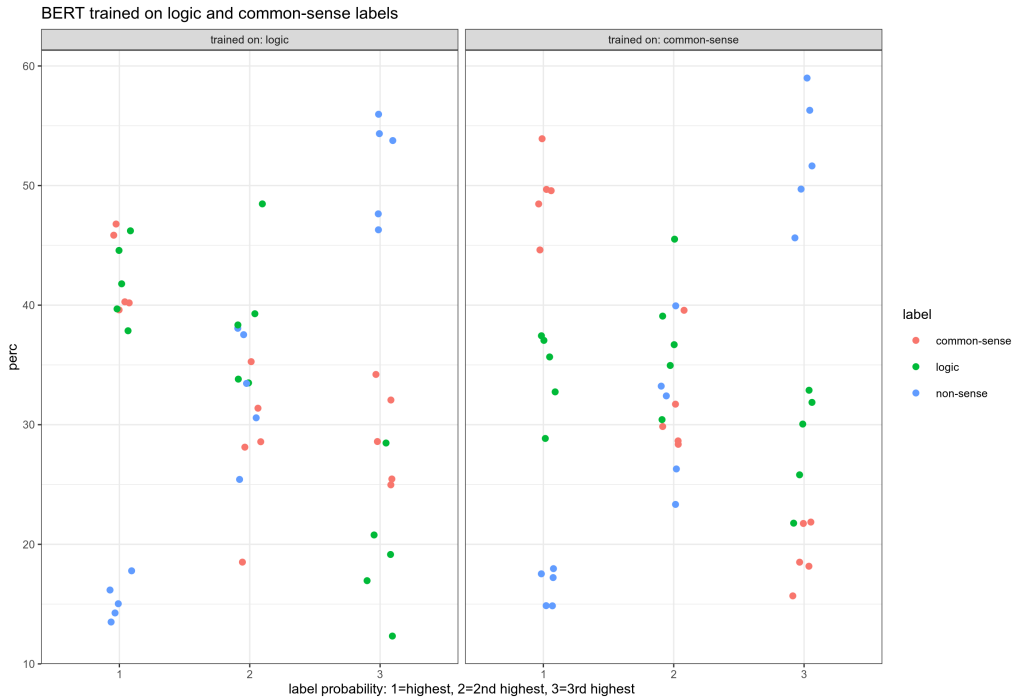

Figure 7
Correlation graph between BERT's predictions and the three annotation stances proposed in this work. On the x-axis we find the three probabilities predicted by the model (because we have 3 inference labels) and on the y-axis we find the percentage of pairs whose logic/common-sense/non-sense label correlates with the model's highest, second highest, and third highest prediction. Each stance is depicted by a different color and the five dots per color represent the running of the model on 5 different seeds.

there is a correlation between the predicted probabilities and the logic and common-sense labels.

To this end, we run the two models on the SICK test set with "distinct" labels and on five different seeds and we compute the correlation graphs shown in Figure 7 for BERT and Figure 8 for RoBERTa. We only test on the subset of pairs that have distinct labels for logic and common-sense because it only makes sense to test on them: If we want to see how the models' predicted probabilities correlate with the different stances of our proposed space, then we need pairs that have different labels for the different stances; otherwise, the picture is not representative. We run the models on five seeds to capture any fluctuation phenomena. The graphs in figures 7 and 8 show the three numbers—the three probabilities (e.g., 70% E, 20% N, 10% C) that the models will predict (since we have three target labels) in the x-axis. These numbers add up to 100 (probability of 1). Particularly, we see that the x-axis shows the label with the highest probability (notation *1*), the label with the second highest probability (notation 2), and the one with the third highest probability (notation *3*). The y-axis shows the percentage of the pairs where the nth probability of the model is the same with the logic/common-sense/non-sense label of our proposed annotation scheme (green, red, and blue notations, respectively). The five dots of each color represent the five seeds we used. The two panels, left and right,

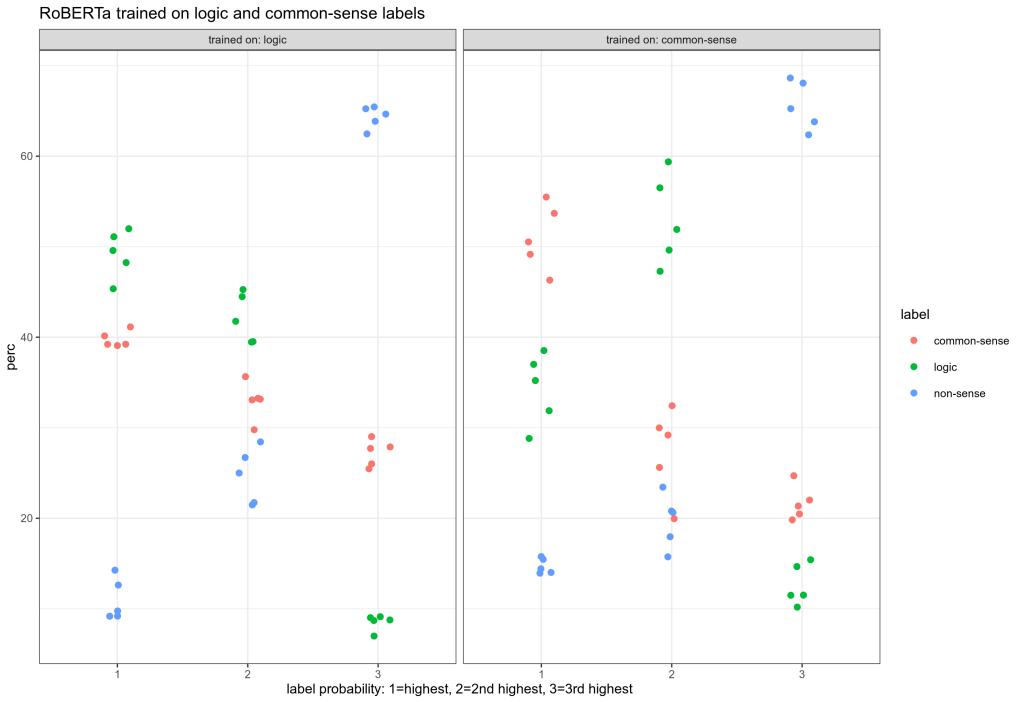

Figure 8
Correlation graph between RoBERTa's predictions and the three annotation stances proposed in this work. On the x-axis we find the three probabilities predicted by the model (because we have 3 inference labels) and on the y-axis we find the percentage of pairs whose logic/common-sense/non-sense label correlates with the model's highest, second highest, and third highest prediction. Each stance is depicted by a different color and the five dots per color represent the running of the model on 5 different seeds.

show the training set we used, that is, whether the logically annotated train set (left) or the common-sense annotated train set (right) was used.

The graphs allow us a number of observations. First, we confirm the results reported in Table 4: RoBERTa is much better than BERT in capturing the different stances introduced in this work. For RoBERTa in Figure 8 we can see clear clusters of the logic/common-sense/non-sense correlations (colors nicely separated), whereas for BERT the data is not separated clearly. This means that BERT struggles to learn these different stances, whether it trained on the logic or the common-sense label. Still, even the performance of RoBERTa is not as high as we would want: The percentages of the two highest probabilities lie at around 50% to 60%, showing that RoBERTa is also not entirely able to differentiate between the different stances. Particularly, we observe the following: When RoBERTa is trained on the logic label (Figure 8, left), the logic percentage is higher than the common-sense percentage for both the highest and the second highest probability. This is also what we would expect: Because it is trained on logic labels, the two most probable predictions should correlate more with the logic labels of the pairs. However, this picture is flipped when training on the common-sense label (Figure 8, right): Here, the common-sense percentage is higher for the highest model probability but for the second highest probability it is the logic percentage that is higher. This means that even RoBERTa is not able to have a consistent performance by

Table 8
Accuracy of pretrained models when fine-tuned on SICK and tested on SNLI and MNLI. The notation *.logic/.common-sense/.original* states which stance the annotator assumed when providing the label.

Testset	**Trainset**					
	original SICK.train		SICK.train.logic		SICK.train.common-sense	
	BERT	**RoBERTA**	**BERT**	**RoBERTA**	**BERT**	**RoBERTA**
SNLI.test	44.8	52.2	43.4	52.3	43.9	53
MNLI.dev.matched	43.9	51.4	42.7	53.7	43.2	57.3
MNLI.dev.mismatched	46	52.7	44.7	55.4	44.9	59.7

predicting the highest probabilities for the stance it has been trained on across the board. This is not surprising considering that these models have not been optimized to capture such different stances, but it also suggests that such models might be able to learn such differences if trained and optimized on suitable data. A further finding for the more efficient RoBERTa is that the non-sensical label correlates with the lowest percentage in the highest and second highest probabilities and with the highest percentage in the third highest probability, no matter the training set. This is indeed sensible because it shows that the non-sensical labels, that is, labels that do not lead to valid inferences, are less likely to be predicted by the model as the first or second probability but rather as the third one. This also indicates more clearly the proposed status of the non-sensical labels.

Testing on Other NLI Datasets. Table 8 shows the performance when training on the original corpus and the newly annotated versions and tested on entirely different data and particularly on SNLI and MNLI. It is clear that, no matter the training set, BERT and RoBERTA struggle to deal with the SNLI and MNLI inferences. This is unexpected for the original SICK training set and expected for the new training sets. For the original train set we would expect a similar performance like the one for the original SICK test set. In particular, this is expected for SNLI, which is supposed to contain the same kind of simple, every-day, caption-based inferences that SICK does. One reason for this strange behavior might be that SNLI and MNLI contain many sentences without a finite verb, for example, pairs like *P: A young boy in a field of flowers carrying a ball. H: boy in field*. Such structures are not found in SICK. It could thus be the case that the model has not learned to deal with such verb-less sentences. For the new re-annotated train sets, the results are expected. The train sets are created based on the ordered sense space annotation proposed in this work and thus the trained model cannot efficiently deal with test sets with "mixed" logical and common-sense annotations, as SNLI and MNLI are.

5.3 Discussion

With these experiments we are able to address our research questions and, in particular we are able to show (a) the inherent difficulty and variability of certain pairs, (b) how current models struggle to efficiently model logic inferences and the common-sense ones (but rather model a mix of them), (c) how the current task formulation does

not allow for the distinct handling of different kinds of inferences and thus inevitably leads models to non-representative performance, and (d) how the proposed ordered sense space annotation can come to the rescue. The ordered sense space proposes to distinguish between strict, logical inferences and common-sense, "people on the street" inferences. In this way, hard and ambiguous pairs are allowed to also have ambiguous, even conflicting, annotations. With such annotations in place, the whole NLI scenery can be transformed. First, the ordered sense space annotation can be exploited during training: Pairs that have the same label for the logic and the common-sense stance are not ambiguous, are thus more reliable for training and can have a stronger learning effect—for example, have higher training weights, than "hard" pairs with distinct labels for logic and common-sense. Such special weighting ought to improve model performance. Second, this kind of annotation can be exploited during evaluation: We can measure performance on pairs with the same and with distinct labels of logic/common-sense. This will give us a better picture of what the models really learn and how they can be improved. Third, the annotation can be used to train models to distinguish between "hard" (ambiguous) and "easy" (unambiguous) pairs. If this is achieved, then models tuned for logic or for common-sense can be combined. In fact, this opens the way for a hybrid NLI direction, where symbolic systems live alongside neural models. If a pair can be identified as "hard," requiring a different label for logic and for common-sense, the label of logic can be given by a symbolic system and the label of common-sense by a neural model, as each of these methods has been shown to excel in strict reasoning and "people in the street" inference, respectively. At the same time, such a "circumscribed" dataset, as Zaenen, Karttunen, and Crouch (2005) call it, can contribute to the higher explainability of current models. We can know whether a model fails at a logical or a common-sense inference and thus, we can also understand more of the inner workings of the black-box. Consequently, we can choose to optimize the parts that we know are failing.

Connection to Previous Formulations of NLI. The proposed ordered sense space is not entirely novel, in that it attempts to combine the two seemingly opposing stands expressed by Zaenen, Karttunen, and Crouch (2005) and Manning (2006) (see Section 2.1). The former argues for a more "controlled" inference task, where the kind of inference and world-knowledge is carefully circumscribed and considered. The latter defends a less precise and more "natural" task, where inferences "in the wild" are captured. The two viewpoints are mostly seen as contradictory, while in reality they are rather complementary. Our proposed approach attempts to combine them into one: annotators are allowed to decide based on their "gut feeling," naturally, but they are also asked to judge whether a more strict view of the world would produce a different inference. Thus, a dataset will naturally end up annotated with some of the types of meta-information that Zaenen, Karttunen, and Crouch (2005) ask for, without at the same time sacrificing the spontaneous, natural, "human on the street" inference mechanism that Manning (2006) and so many others consider important. Except for these works, our proposed approach is also partly parallel to the proposals by Pavlick and Callison-Burch (2016): The researchers show that a distribution over the inference labels would be a better learning target for NLI models because it captures more faithfully the inherent disagreement of human inferences. In our approach we also capture this inherent disagreement and distribution of labels by implementing a continuum in the middle of the ordered sense space. But we do more than that: We practically *meta-annotate* this label distribution as common-sense reasoning and explicitly distinguish it from logical and non-sensical reasoning, which should be rigid and uniform. Overall, our approach

can be situated between older and more recent proposals, bridging the gap between their different desiderata and goals.

But Is the Sense Space Real? We have presented our ordered sense space in terms of humans and the stances they assume when doing Natural Language Inference in ordinary life. If we are right, then this space already shows up in the labels attached by human annotators, the labels that determine the "gold standards" of datasets. But can we find anything like this in the performance of our models? It turns out that we can. Our re-annotation provides data *along with the sense stance that was used in determining the label of each pair*. We thus can train and test on differing subsets. We have shown that there is a difference. Table 5 shows that pairs with distinct labels obtained by either the logic or the common-sense stance are difficult for models. The correlations offered in Figures 7 and 8 also show to what extent the distribution space currently predicted by the models can reflect the different sense stances.

Cost for Our Annotation. We estimate that using our ordered sense space annotation will take 1.5 times the amount of time as the single-label annotation, but we believe this is worthwhile considering the more precise and representative of human inference annotation we obtain. Despite this longer time, the annotation is definitely feasible as a crowdsourcing task as it is supposed to be self-explanatory. Also, making the distinction between strict, logical inference and the more casual, common-sense reasoning forces the annotator to think harder about the justifications for their annotation. Furthermore, it should be noted that the required "space thinking" can easily be picked up while annotating, so that later annotations are faster than the first ones. Last but not least, we think our method is in line with NLI datasets asking annotators to provide explanations for their annotations (Camburu et al. 2018; Nie et al. 2020), a step toward providing more information behind the annotators' decision and building models that capture the nuance of human annotated data (Prabhakaran, Davani, and Díaz 2021).

6. A Small-Scale Crowd Annotation with MNLI

In order to see whether our method could be applied to a crowd annotation setting, we conducted a small annotation experiment on the Amazon MTurk platform,[13] with 100 examples from MNLI (Williams, Nangia, and Bowman 2018). These examples are randomly sampled from the chaosNLI (Nie, Zhou, and Bansal 2020) dataset because they are the ones where the 5 annotators did not reach a unanimous decision in the original SNLI/MNLI annotation.

Specifically, we adapted our scheme slightly and asked the crowd annotators to give two inference labels for a given pair, based on two stances: one as a judge in court (corresponding to the strict logical annotation), and another as a person on the street (roughly equivalent to the common-sense annotation). We restricted our annotators to those based in the US. Each pair was annotated by 10 Turkers, and one HIT contained 22 pairs, with 20 pairs from the MNLI portion of chaosNLI and another two as a catch trial. If the annotator answered the catch trial wrong, then their data would be excluded in the final analysis. They were compensated $4 USD for annotating one HIT.[14] The average completion time for one HIT was about 18 minutes.

13 https://www.mturk.com/.

14 Full instructions can be found here: https://huhailinguist.github.io/stuff/Instructions.pdf.

Table 9
Distribution of pairs with different labels from the crowd annotation experiment. Note: 'EN' means that there is a tie between "E" and "N".

as a judge: majority label	as a person on the street: majority label	number of pairs
N	E	38
N	N	21
C	C	14
E	E	10
EN	E	6
NC	C	3
N	EN	3
N	C	2
N	EC	1
N	NC	1
ENC	E	1

In the total of five HITs (20 pairs per HIT) we distributed, eight annotators (out of 50) were excluded because their answers to the catch trials were wrong.

For each stance, we took the majority vote of the labels. Note that there might be a tie between two or even three labels. The results are summarized in Table 9.[15] From the first row, we clearly see that annotators are able to distinguish between these two stances: 38 pairs are judged to be "neutral" if they take a strict stance as if they were a judge in the court but "entailment" if they took a loose criterion as a person on the street. The next three rows in the table are pairs that have the same labels for both stances, which constitute 45 pairs in total. Then we have a couple of pairs that are labeled "neutral" as a judge, but "contradiction" to a person on the street.

Using data from chaosNLI allows us to compare the two-stance annotations we obtained with the 100 annotations the chaosNLI authors obtained under the single-label scheme. For instance, consider the following pair:

premise:	Most of Slate will not be published next week, the third and last of our traditional summer weeks off.
hypothesis:	Slate won't be published much this summer.

This received 62 "entailment" labels, 31 "neutral" labels, and 7 "contradiction" labels in chaosNLI. In our annotation, it was labeled as "neutral" from the stance as a judge, and "entailment" from the person on the street stance. This indicates that our two annotation schemes converge on this pair, and what is more interesting is that our two-label scheme offers more information: Those who annotate it as "entailment" are being loose in their judgments, while those giving it "neutral" are being more strict. This result makes sense as how "much" should be considered as "published much" is a rather subjective decision, which may vary from annotator to annotator, giving rise to the "neutral" label if one were to be strict on what counts as an "entailment."

15 All annotations can be found at `https://github.com/huhailinguist/curing-SICK`.

Overall, we take our small crowd annotation experiment to indicate that non-logicians or non-linguists are capable of making the distinction between the two stances, and it is possible to apply our annotation scheme in the future for the construction of other NLI datasets.

7. Conclusion and Future Work

In this paper we have conducted an extensive study of the SICK corpus and have shed light on the aspects of NLI that make it so error-prone and uncertain. To tackle these challenges, this work proposes a new annotation method, which distances itself from the single label approach and the graded certainty distribution task. Instead, we have proposed an ordered sense space annotation that is able to combine the typical human inference mechanism, which is natural, spontaneous, and loose, with a more guided, precise, and logical mechanism that is required in some tasks. The proposed ordered sense space annotation allows for each inference problem to be solved separately based on strict logic and common-sense, not only accepting conflicting annotations but also being able to give explanations for them. The efficiency of the proposed annotation has been shown through specific corpus examples, which are split into formal categories, and also through thorough experiments with transformer-based neural models. We are able to show marked differences in the performance of pretrained models when tested on pairs with inherent ambiguity and variability, exposing the models' ability to distinguish between logical and common-sense inference. With this, we are also able to propose ways in which current models and the whole NLI landscape could benefit from the ordered sense space annotation. By achieving these goals, we have also provided a new version of the SICK corpus, re-annotated and reliable, to be used for training and testing. We have offered a taxonomy of annotation issues and guidelines. We have also shown that out method can be scaled up to crowd-sourcing annotation, on a different NLI dataset, MNLI. Overall, this work contributes to the current NLI field by (a) providing a reliable corpus, (b) providing a taxonomy of challenging NLI annotations, (c) proposing a new annotation scheme that tackles these challenges, (d) showing the practical benefits of the annotation scheme when evaluated on state-of-the-art transformer models, and (e) providing initial results of crowdsourcing using the proposed scheme. The final corpus with the ordered sense space annotations, the labels of the symbolic systems, as well as the crowd-annotated examples from MNLI are available at `https://github.com/huhailinguist/curing-SICK`.

Going forward, we would like to extend this work. First, we would like to explore ways to make better use of the common-sense annotation. Currently and due to the small number of relevant occurrences (see Section 4.3), the common-sense annotation is interpreted as a distinct label for the statistics and the model experiments. However, it would be interesting to investigate how the gradient scale available within the common-sense continuum can be exploited for the training and testing of the models, possibly along the lines of the proposals by Pavlick and Kwiatkowski (2019). For example, we could explore whether the models can capture the tendency of the common-sense label, that is, whether it is rather toward the logical end or the non-sensical one. In this way, this future direction could promote the dialog between researchers who are skeptical of the current NLI task formulation. Finally, future work should attempt to devise a new corpus based on the proposed ordered sense space annotation and train suitable models. This effort will certainly highlight open issues of the approach, but will also open the way for the use of such corpora in the training and testing of neural models, contributing to improved, well-formed NLI.

Appendix A. Annotation Guidelines

In this section, we lay out our annotation guidelines, as those were re-formed based on the proposed sense annotation. Most importantly, we categorize the difficult cases we encountered in our pilot annotation. We also highlight the cases where more than one distinctive label may be needed.

Semantically Absurd Pairs

Description: One of the sentences is semantically absurd.
Prototypical Example: *P: A motorcycle rider is standing up on the seat of a white motorcycle. H: A motorcycle is riding standing up on the seat of the vehicle.*
Solution: Mark it as semantically absurd (AB).

Ungrammatical

Description: There is at least one grammatical error in one of the sentences.
Prototypical Example: *P: A man in a car is pulling up beside a woman that is who along a road. H: A man in a car is not pulling up beside a woman who is walking along a road.*
Solution: Mark them as ungrammatical (UG).

Aspect/Tense

Description: Different tenses and aspects are involved. The inference label may be different depending on whether we take the tenses and aspects literally or ignore them.
Prototypical Example: *P: The adults and children are not gathered near an outdoor seating arrangement. H: The adults and children are being gathered near an outdoor seating arrangement.*
Solution: Assume the present and progressive. Thus, the above example should be a contradiction.

Phrasal Verbs - Prepositions

Description: There is a phrasal verb involved or part of it and depending on what we take it to mean, the inference label could change. Or there is some preposition that seems to be altering the meaning of the sentence.
Prototypical Example: *P: A woman is picking a can. H: A woman is taking a can.*
Solution: Assign two labels, based on logic and common-sense, as see fit.

Idiomatic Phrases

Description: An idiomatic phrase is involved. Depending on how the phrase is interpreted, the label may be different.
Prototypical Example: The couple is walking down the aisle. The couple is walking up the aisle.
Solution: Assign two labels, based on logic and common-sense, as see fit.

Looseness

Description: It is not clear how strict we should be with the meaning of the word. We can decide to be more loose to match the context.

Prototypical Example: *P: A surfer is surfing a big wave. H: A surfer is surfing a huge wave.*
Another example: *P: A small guinea pig is gnawing and eating a piece of carrot on the floor. H: A guinea pig is devouring a carrot.* (Note: some may argue that guinea pigs always devour; they cannot eat slowly, or in a non-devouring manner.)
Yet another example: *P: A woman is cutting broccoli. H: A woman is slicing vegetables*
Solution: Assign two labels, based on logic and common-sense, as see fit.

Complete Opposite Subjects, but Same Predicate

Description: The subjects are dictionary antonyms, but the rest of the event is the same.
Prototypical Example: *P: A few men are dancing. H: Some women are dancing.*
Solution: If subjects are antonyms, then label the pair as C. If not, then check the two words to see what relation they have (E or N).

Complete Opposite Predicates, but Same Subject

Description: The predicates are antonyms, but the agent is the same.
Prototypical Example: *P: The man is fasting. H: The man is eating.*
Solution: Assign two labels, based on logic and common-sense, as see fit.

Examples Involving Gender, Age, or Title

Description: It is not clear whether words like *man* and *woman* should be taken to refer to the gender or to the age aspect: e.g., girl - woman, boy - man, woman - lady.
Prototypical Example: *P: A girl is eating a cupcake. H: A woman is not eating a cupcake.*
Solution: Assign two labels, based on logic and common-sense, as see fit.

Privative Adjectives

Description: Examples involving a privative adjective, e.g., fake, cartoon, animated, etc.
Prototypical Example: *P: An animated airplane is landing. H: The plane is landing.*
Solution: Assign two labels, based on logic and common-sense, as see fit.

Different Senses

Description: There are different senses available for some words and depending on which one we choose to interpret the sentences, the label could be different.
Prototypical Example: *P: A person is folding a sheet. H: A person is folding a piece of paper.*
Another Example: *P: Three men and a woman are standing still for a picture. H: A woman and three men are posing for a photo.*
Solution: Assign two labels, based on logic and common-sense, as see fit.

Ambiguous Sentences

Description: The premise or hypothesis is ambiguous.
Example: *P: Two men are fighting in a cattle pen. H: Two men are fighting in a pen for cattle.*
"Two men are fighting in a pen for cattle" has PP attachment ambiguity: it could be a cattle pen, or fighting to win the cattle.
Solution: Label with the label that seems most natural to your common-sense.

World Knowledge

Description: Words in the premise or hypothesis require world knowledge for interpretation. For instance, different annotators may have different interpretations for the inference relations of the following word/phrase pairs:

- football vs. soccer
- beach by the lake vs. beach by the sea
- frying vs. cooking breaded pork chop
- flying vs. travelling

Solution: Label with the label that seems most natural to your world-knowledge.

Coreference

Description/Discussion: We try to make the sentences coreferent based on the subject or the verb (event).
Examples and solutions:

- *P: A woman is walking. H: A man is walking*: Verbs (events) can be made coreferent, so C.
- *P: A woman is walking. H: A woman is dancing*: Subjects can be made coreferent, so C.
- *P: The dog is eating a bone. H: A cat is playing with a bone*: The bone is not enough to make it coreferent, so N.
- *P: The man is walking. H: The dog is eating*: Nothing can be made coreferent, so N. Note that this is different from the guidelines of SNLI (Bowman et al. 2015), where irrelevant pairs are annotated as C.

Appendix B: Precision, Recall, F1 Score for Symbolic Systems

We present the precision, recall, and F1-score for logic systems on SICK.test in Table 10.

Table 10
Precision, recall, and F1-score for symbolic systems.

	ccg2lambda1	ccg2lambda12	LangPro	MonaLog	GKR4NLI
original SICK.test					
C (P/R)	99.2|65.2	98.9|64.0	97.3|65.5	81.2|64.5	74.6|80.3
E (P/R)	96.9|55.8	96.9|54.2	97.4|55.2	97.8|41.1	76.5|60.5
N (P/R)	76.1|99.1	75.4|99.0	76.0|99.0	71.2|95.7	78.3|85.0
weighted F1	80.6	79.9	80.4	73.4	76.9
SICK.test.logic					
C (P/R)	99.8|44.9	99.8|44.2	98.5|45.2	98.9|53.5	89.4|65.2
E (P/R)	97.0|60.7	98.5|59.7	97.4|59.8	99.3|45.2	77.6|65.5
N (P/R)	69.9|99.4	69.4|99.6	69.8|99.3	67.7|99.6	77.6|92.6
weighted F1	75.8	75.4	75.6	73.3	78.9
SICK.test.common-sense					
C (P/R)	99.6|44.0	99.6|43.2	98.1|44.2	99.1|52.6	90.0|64.4
E (P/R)	97.7|58.0	98.2|56.6	97.9|57.1	99.0|42.8	78.9|63.2
N (P/R)	67.6|99.5	67.0|99.5	67.4|99.5	65.4|99.6	75.4|93.2
weighted F1	74.1	73.4	73.8	71.3	77.9

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. We thank Livy Real for her help at the initial stage of this project. This article was partly supported by funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within projects BU 1806/10-2 "Questions Visualized" of the FOR2111. This article was partly supported by the Humanities and Social Sciences Grant from the Chinese Ministry of Education (No. 22YJC740020) awarded to Hai Hu. The work of Lawrence S. Moss was supported by grant #586136 from the Simons Foundation.

References

Abzianidze, Lasha. 2014. Towards a wide-coverage tableau method for natural logic. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2014 Workshops, LENLS, JURISIN, and GABA, Kanagawa, Japan, October 27–28, 2014, Revised Selected Papers*, pages 66–82. https://doi.org/10.1007/978-3-662-48119-6_6

Abzianidze, Lasha. 2015. A tableau prover for natural logic and language. In *Proceedings of EMNLP*, pages 2492–2502. https://doi.org/10.18653/v1/D15-1296

Abzianidze, Lasha. 2016. *A Natural Proof System for Natural Language*. Ph.D. thesis, Tilburg University.

Abzianidze, Lasha. 2017. LangPro: Natural language theorem prover. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 115–120. https://doi.org/10.18653/v1/D17-2020

Bar-Haim, Roy, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop*, pages 1–9.

Bentivogli, Luisa, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2010. The sixth PASCAL recognizing textual entailment challenge. In *Proceedings of the 10th Text Analysis Conference (TAC)*, 18 pages.

Bentivogli, Luisa, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the 11th Text Analysis Conference (TAC)*, 16 pages.

Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge.

In *Proceedings of the 9th Text Analysis Conference (TAC)*, 15 pages.

Bhagavatula, Chandra, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020*, 18 pages.

Bobrow, Daniel G., Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria, de Paiva, Charlotte Price, and Annie, Zaenen. 2007. PARC's bridge and question answering system. In *Proceedings of the Grammar Engineering Across Frameworks Workshop (GEAF 2007)* pages 46–66.

Bowman, Samuel R., 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.

Camburu, Oana Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with natural language explanations. In *NeurIPS*, pages 9539–9549.

Chen, Tongfei, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain Natural Language Inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779.

Chierchia, G. and S. McConnell-Ginet. 2001. *Meaning and Grammar: An Introduction to Semantics*, 2 edition. MIT Press.

Clark, Stephen and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552. https://doi.org/10.1162/coli.2007.33.4.493

Cooper, Robin, Stergios Chatzikyriakidis, and Simon Dobnik. 2016. Testing the FraCaS test suite. In *Proceedings of the Logic and Engineering of Natural Language Semantics Workshop*.

Crouch, Richard, Lauri Karttunen, and Annie Zaenen. 2006. Circumscribing is not excluding: A response to Manning. Unpublished manuscript. http://www2.parc.com/istl/members/karttune/publications/reply-tomanning.pdf.

Crouch, Richard and Tracy Holloway King. 2007. Systems and methods for detecting entailment and contradiction. US Patent 7,313,515.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the Machine Learning Challenges Workshop*, pages 177–190.

Dasgupta, Ishita, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. Evaluating compositionality in sentence embeddings. *CoRR*, abs/1802.04302.

de Marneffe, Marie Catherine, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047.

Devlin, Jacob, 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.

Fitting, Melvin. 1990. *First-order Logic and Automated Theorem Proving*. Springer-Verlag.

Geva, Mor, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166. https://doi.org/10.18653/v1/D19-1107

Giampiccolo, Danilo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2008. The fourth PASCAL recognizing textual entailment challenge. In *Proceedings of the 8th Text Analysis Conference (TAC)*, 8 pages. https://doi.org/10.3115/1654536.1654538

Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pages 1–9. https://doi.org/10.3115/1654536.1654538

Glickman, Oren, Ido Dagan, and Moshe Koppel. 2005. A probabilistic classification approach for lexical textual entailment. *Proceedings of the National Conference on Artificial Intelligence*, 3:1050–1055.

Glockner, Max, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655. https://doi.org/10.18653/v1/P18-2103

Gorman, Kyle and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in Natural Language Inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. https://doi.org/10.18653/v1/N18-2017

Horn, Laurence. 1989. *A Natural History of Negation*. University of Chicago Press.

Hu, Hai, Qi Chen, and Larry Moss. 2019. Natural Language Inference with monotonicity. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 8–15.

Hu, Hai, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S. Moss, and Sandra Kuebler. 2020. MonaLog: A lightweight system for Natural Language Inference based on monotonicity. In *Proceedings of SCiL*, pages 334–344.

Hu, Hai and Lawrence S. Moss. 2018. Polarity computations in flexible categorial grammar. In *Proceedings of *SEM*, pages 124–129. https://doi.org/10.18653/v1/S18-2015

Kalouli, Aikaterini Lida. 2021. *Hy-NLI: A Hybrid system for state-of-the-art Natural Language Inference*. Ph.D. thesis, Universität Konstanz, Konstanz.

Kalouli, Aikaterini Lida, Annebeth Buis, Livy Real, Martha Palmer, and Valeria dePaiva. 2019. Explaining simple Natural Language Inference. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143.

Kalouli, Aikaterini Lida and Richard Crouch. 2018. GKR: The Graphical Knowledge Representation for semantic parsing. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 27–37. https://doi.org/10.18653/v1/W18-1304

Kalouli, Aikaterini Lida, Richard Crouch, and Valeria de Paiva. 2020. Hy-NLI: A hybrid system for Natural Language Inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING '20, pages 5235–5249. https://doi.org/10.18653/v1/2020.coling-main.459

Kalouli, Aikaterini Lida, Livy Real, Annebeth Buis, Martha Palmer, and Valeria de Paiva. 2021. Annotation difficulties in Natural Language Inference. In *Proceedings of the Symposium in Information and Human Language Technology*, pages 247–254. https://doi.org/10.5753/stil.2021.17804

Kalouli, Aikaterini Lida, Livy Real, and Valeria de Paiva. 2017a. Correcting contradictions. In *Proceedings of Computing Natural Language Inference (CONLI) Workshop*, 6 pages.

Kalouli, Aikaterini Lida, Livy Real, and Valeria de Paiva. 2017b. Textual inference: Getting logic from humans. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, 7 pages.

Kalouli, Aikaterini Lida, Livy Real, and Valeria de Paiva. 2018. WordNet for "easy" textual inferences. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 35–42. https://doi.org/10.1609/aaai.v32i1.12022

Khot, Tushar, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A textual entailment dataset from science question answering. In *AAAI*, pages 5189–5197.

Lewis, Mike and Mark Steedman. 2014. A* CCG parsing with a supertag-factored model. In *Proceedings of EMNLP*, pages 990–1000. https://doi.org/10.3115/v1/D14-1107

Liu, Xiaodong, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

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

MacCartney, Bill. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University. AAI3364139.

Manning, Christopher D. 2006. Local textual inference: It's hard to circumscribe, but

you know it when you see it—and NLP needs it. https://nlp.stanford.edu/manning/papers/TextualInference.pdf.

Marelli, Marco, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8. https://doi.org/10.3115/v1/S14-2001

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. 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.

Mates, Benson. 1972. *Elementary Logic*. Oxford University Press.

McCoy, Tom, 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. https://doi.org/10.18653/v1/P19-1334

Naik, Aakanksha, 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.

Nie, Yixin, Yicheng Wang, and Mohit Bansal. 2018. Analyzing compositionality-sensitivity of NLI models. *CoRR*, abs/1811.07033:6867–6874. https://doi.org/10.1609/aaai.v33i01.33016867

Nie, Yixin, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of ACL*, pages 4885–4901. https://doi.org/10.18653/v1/2020.acl-main.441

Nie, Yixin, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on Natural Language Inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143. https://doi.org/10.18653/v1/2020.emnlp-main.734

Partee, Barbara H. 2010. Privative adjectives: Subsective plus coercion. *Presuppositions and Discourse: Essays Offered to Hans Kamp*. Brill, pages 273–285. https://doi.org/10.1163/9789004253162_011

Pavlick, Ellie and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173. https://doi.org/10.18653/v1/P16-1204

Pavlick, Ellie and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Pilault, Jonathan, Amine Elhattami, and Christopher Pal. 2020. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. arXiv preprint arXiv:2009.09139v1.

Poesio, Massimo and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83. https://doi.org/10.3115/1608829.1608840

Poliak, Adam, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191. https://doi.org/10.18653/v1/S18-2023

Prabhakaran, Vinodkumar, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138. https://doi.org/10.18653/v1/2021.law-1.14

Richardson, Kyle, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing Natural Language Inference models through semantic fragments. In the *Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, the *Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020*, the *Tenth*

AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, pages 8713–8721. https://doi.org/10.1609/aaai.v34i05.6397

Sánchez-Valencia, Victor. 1991. *Studies on Natural Logic and Categorial Grammar*. Ph.D. thesis, University of Amsterdam.

van Benthem, Johan. 2008. A brief history of natural logic. In M. Chakraborty, B. Lowe, M. Nath Mitra, and S. Sarukki, editors, *Logic, Navya-Nyaya and Applications: Homage to Bimal Matilal*. College Publications.

van der Goot, Rob. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494. https://doi.org/10.18653/v1/2021.emnlp-main.368

Versley, Yannick. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6:333–353.

Williams, Adina, 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. https://doi.org/10.18653/v1/N18-1101

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yanaka, Hitomi, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117. https://doi.org/10.18653/v1/2020.acl-main.543

Yanaka, Hitomi, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of BlackboxNLP*, pages 31–40. https://doi.org/10.18653/v1/W19-4804

Yanaka, Hitomi, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of *SEM*, pages 250–255. https://doi.org/10.18653/v1/S19-1027

Yanaka, Hitomi, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. 2018. Acquisition of phrase correspondences using natural deduction proofs. In *Proceedings of NAACL*, pages 756–766.

Zaenen, Annie, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36. https://doi.org/10.3115/1631862.1631868

Zhang, Sheng, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395. https://doi.org/10.1162/tacl_a_00068