false-false, 2,262 (20%) *true-false*, and 1,131 (10%) *random*. We include all possible 5,655 *true-true* combinations of 30 true sentences for each of the 13 source sentences. For *false-false*, *true-false*, and *random* we downsample the full set of pairs to obtain the desired number, keeping an equal number of samples per source sentence. We chose this distribution because we are mainly interested in paraphrases and entailment, as well as their relation to specificity. We hypothesize that pairs of sentences that are both true have the highest potential to contain these relations.

From the 11,310 pairs, we randomly selected 520 (5%) for annotation, with the same 50-20-20-10 distribution as the full corpus. We select an equal number of pairs from each source sentence. We hypothesize that length strongly correlates with specificity, as there is potentially more information in a longer sentence that in a shorter one. Hence, for half of the pairs, we made sure that the difference in length between the two sentences is not more than 1 token.

7.3.3 Relation Annotation

We annotate all the relations in the corpus of 520 sentence pairs using Amazon Turk. We select 10 crowdworkers per task, as this gives us the possibility to measure how well the tasks has been understood overall, but especially how easy or difficult individual pairs are in the annotation of a specific relation. In the SICK corpus, the same platform and number of annotators were used.

We chose to annotate the relations separately to avoid biasing the crowdworkers who might learn heuristic shortcuts when seeing the same relations together too often. We launched the tasks consecutively to have the annotations as independent as possible. This differs from the SICK corpus annotation setting, where entailment, contradiction, and semantic similarity were annotated together.

The complex nature of the meaning relations makes it difficult to come up with a precise and widely accepted definition and annotation instructions for each of them. This problem has already been emphasized in previous annotation tasks and theoretical settings [Bhagat and Hovy, 2013]. The standard approach in most of the existing paraphrasing and entailment datasets is to use a more generic and less strict definitions. For example, pairs annotated as "paraphrases" in MRPC [Dolan et al., 2004] can have "obvious differences in information content". This "relatively loose definition of semantic equivalence" is adopted in most empirically oriented paraphrasing corpora.

We take the same approach towards the task of annotating semantic relations: we provide the annotators with simplified guidelines, as well as with few positive and negative examples. In this way, we believe that annotation is more generic, reproducible, and applicable to any kind of data. It also relies more on the intuitions of a competent speaker than on understanding complex linguistic concepts.

Prior to the full annotation, we performed several pilot studies on a sample of the corpus in order to improve instructions and examples given to the annotators. In the following, we will shortly outline the instructions for each task.

Paraphrasing In Paraphrasing (PP), we ask the crowdworkers whether the two sentences have approximately the same meaning or not, which is similar to the definition of Bhagat and Hovy [2013] and De Beaugrande and Dressler [1981].

Textual Entailment In Textual Entailment (TE), we ask whether the first sentence makes the second sentence true. Similar to RTE Tasks [Dagan et al., 2006] - [Bentivogli et al., 2011], we only annotate for forward entailment (FTE). Hence, we use the pairs twice: in the order we ask for all other tasks and in reversed order, to get the entailment for both directions. Backward Entailment is referred to as *BTE*. If a pair contains only backward or forward entailment, it is uni-directional (UTE). If a pair contains both forward and backward entailment, it is bi-directional (BiTE). Our annotation instructions and the way we interpret directionality is similar to other crowdworking tasks for textual entailment [Marelli et al., 2014, Bowman et al., 2015].

Contradiction In Contradiction (Cont), we ask the annotators whether the sentences contradict each other. Here, our instructions are different from the typical approach in RTE [Dagan et al., 2006], where contradiction is often understood as the absence of entailment.

Specificity In Specificity (Spec), we ask whether the first sentence is more specific than the second. To annotate specificity in a comparative way is new [4]. Like in textual entailment, we pose the task only in one direction. If the originally first sentence is more specific, it is forward specificity (FSpec), whereas if the originally second sentence is more specific than the first, it is backward specificity (BSpec).

Semantic Similarity For semantic similarity (Sim), we do not only ask whether the pair is related, but rate the similarity on a scale 0-5. Unlike previous studies [Agirre et al., 2014], we decided not to provide explicit definitions for every point on the scale.

Annotation Quality To ensure the quality of the annotations, we include 10 control pairs, which are hand-picked and slightly modified pairs from the original corpus, in each task.[5] We discard workers who perform bad on the control pairs. [6]

[4] Louis and Nenkova [2012] labelled individual sentences as *specific*, *general*, or *cannot decide*.

[5] The control pairs are also available online at `https://github.com/MeDarina/meaning_relations_interaction`

[6] Only 2 annotators were discarded across all tasks. To have an equal number of annotations for each task, we re-annotated these cases with other crowdworkers.

7.3.4 Final Corpus

For each sentence pair, we get 10 annotations for each relation, namely paraphrasing, entailment, contradiction, specificity, and semantic similarity. Each sentence pair is assigned a binary label for each relation, except for similarity. We decide that if the majority (at least 60% of the annotators) voted for a relation, it gets the label for this relation.

Table 7.8 shows exemplary annotation outputs of sentence pairs taken from our corpus. For instance, sentence pair #4 contains two relations: forward entailment and forward specificity. This means that it has uni-directional entailment and the first sentence is more specific than the second. The semantic similarity of this pair is 2.7.

Inter-Annotator Agreement We evaluate the agreement on each task separately. For semantic similarity, we determine the average similarity score and the standard deviation for each pair. We also calculate the Pearson correlation between each annotator and the average score for their pairs. We report the average correlation, as suggested by SemEval [Agirre et al., 2014] and SICK.

For all nominal classification tasks we determine the majority vote and calculate the % of agreement between the annotators. This is the same measure used in the SICK corpus. Following the approach used with semantic similarity, we also calculated Cohen's $kappa$ between each annotator and the majority vote for their pairs. We report the average $kappa$ for each task.[7]

Table 7.2 Inter-annotator agreement for binary relations
✓denotes a relation being there
✗denotes a relation not being there

	%	κ	%✓	%✗	control
PP	.87	.67	.83	.90	.98
TE	.83	.61	.75	.89	.89
Cont	.94	.71	.84	.95	.95
Spec	.80	.56	.81	.82	.89

Table 7.2 shows the overall inter-annotator agreement for the binary tasks. We report: 1) the average %-agreement for the whole corpus; 2) the average κ score; 3) the average %-agreement for the pairs where the majority label is *"yes"*; 4) the average %-agreement for the pairs where the majority label is *"no"*; 5) the average

[7] We are aware that κ does not fit the restrictions of our task very well and also that it is usually not averaged. However, we wanted to report a chance corrected measure, which is non-trivial in a crowd-sourcing setting, where each pair is annotated by a different set of annotators.