% agreement between the annotators and the expert-provided "control labels" on the control questions.

The overall agreement for all tasks is between .80 - .94, which is quite good given the difficulty of the tasks. Contradiction has the highest agreement with .94. It is followed by the paraphrase relation, which has an agreement of .87. The agreements of the entailment and specificity relations are slightly lower, which reflects that the tasks are more complex. SICK report agreement of .84 on entailment, which is consistent with our result.

The agreement is higher on the control questions than on the rest of the corpus. We consider it the upper boundary of agreement. The agreement on the individual binary classes shows that, except for the specificity relation, annotators have a higher agreement on the absence of relation.

Table 7.3 Distribution of Inter-annotator agreement

	50%	60%	70%	80%	90%	100%
PP	.11	.12	.13	.20	.24	.20
TE	.17	.19	.17	.16	.19	.10
Cont	.04	.07	.18	.23	.23	.25
Spec	.22	.18	.21	.13	.13	.12

Table 7.3 shows the distribution of agreement for the different relations. We take all pairs for which at least 50% of the annotators found the relation and shows what percentage of these pairs have inter-annotator agreement of 50%, 60%, 70%, 80%, 90%, and 100%. We can observe that, with the exception of contradiction, the distribution of agreement is relatively equal. For our initial corpus analysis, we discarded the pairs with 50% agreement and we only considered pairs where the majority (60% or more) of the annotators voted for the relation. However, the choice of agreement threshold an empirical question and the threshold can be adjusted based on particular objectives and research needs.

The average standard deviation for semantic similarity is 1.05. SICK report average deviation of .76, which is comparable to our result, considering that they use a 5 point scale (1-5), and we use a 6 point one (0-5). Pearson's r between annotators and the average similarity score is 0.69 which is statistically significant at $\alpha = 0.05$.

Distribution of Meaning Relations Table 7.4 shows that all meaning relations are represented in our dataset. We have 160 paraphrase pairs, 195 textual entailment pairs, 68 contradiction pairs, and 381 specificity pairs. There is only a small number of contradictions, but this was already anticipated by the different pairings. The distribution is similar to Marelli et al. [2014] in that the set is

Table 7.4 Distribution of meaning relations within different pair generation patterns

	all	T/T	F/F	T/F	rand.
PP	31%	49%	27%	2%	6%
TE	38%	60%	36%	2%	2%
Cont.	13%	0%	10 %	56%	0%
Spec	73%	79%	72%	66%	63%
∅Sim	2.27	2.90	2.39	1.32	0.77

slightly leaning towards entailment[8]. Furthermore, the distribution of uni- and bi-directional entailment with our and the SICK corpus are similar: they are nearly equally represented.[9]

Distribution of Meaning Relations with Different Generation Pairings Table 7.4 shows the distribution of meaning relations and the average similarity score in the differently generated sentence pairings. In the true/true pairs, we have the highest percentage of paraphrase (49%), entailment (60%), and specificity (79%). In the false/false pairs, all relations of interest are present: paraphrases (27%), entailment (36%), and specificity (72%). Unlike in true/true pairs, false/false ones include contradictions (10%). True/false pairs contain the highest percentage of contradiction (85%). There were also few entailment and paraphrase relations in true/false pairs. In the random pairs, there were only few relations of any kind. The proportion of specificity is high in all pairs.

This different distribution of phenomena based on the source sentences can be used in further corpus creation when determining the best way to combine sentences in pairs. In our corpus, the balanced distribution of phenomena we obtain justifies our pairing choice of 50-20-20-10.

Lexical Overlap within Sentence Pairs As discussed by Joao et al. [2007], a potential flaw of most existing relation corpora is the high lexical overlap between the pairs. They show that simple lexical overlap metrics pose a competitive baseline for paraphrase identification. Due to our creation procedure, we reduce this problem. In Table 7.5, we quantified it by calculating unigram and bigram BLEU score between the two texts in each pair for our corpus, MRPC and SNLI, which

[8] As opposed to contradiction. However, as contradiction and entailment were annotated exclusively, it is not directly comparable.

[9] In SICK 53% of the entailment is uni-directional and 46% are bi-directional, whereas we have 44% uni-directional and 55% bi-directional.

are the two most used corpora for paraphrasing and textual entailment. The BLEU score is much lower for our corpus that for MRPC and SNLI.

Table 7.5 Comparison of BLEU scores between the sentence pairs in different corpora

	MRPC	SNLI	Our corpus
unigram	61	24	18
bigram	50	12	6

Relations and Negation Our corpus also contains multiple instances of relations that involve negations and also double negations. Those examples could pose difficulties to automatic systems and could be of interest to researchers that study the interaction between inference and negation. Pairs #1, #2, and #9 in Table 7.8 are examples for pairs containing negation in our corpus.

7.4 Interactions between Relations

We analyze the interactions between the relations in our corpus in two ways. First, we calculate the correlation between the binary relations and the interaction between them and similarity. Second, we analyze the overlap between the different binary relations and discuss interesting examples.

7.4.1 Correlations between Relations

We calculate correlations between the binary relations using the Pearson correlation. For the correlations of the binary relations with semantic similarity, we discuss the average similarity and the similarity score scales of each binary relation.

7.4.1.1 Correlation of Binary Meaning Relations

In Table 7.6, we show the Pearson correlation between the meaning relations. For entailment, we show the correlation for uni-directional (UTE), bi-directional (BTE), and any-directional (TE).

Paraphrases and any-directional entailment are highly similar with a correlation of .75. Paraphrases have a much higher correlation with bi-directional entailment (.70) than with uni-directional entailment (.20). Prototypical examples of