can be seen as paraphrasing. Furthermore, according to Androutsopoulos and Malakasiotis [2010] both entailment and paraphrasing are intended to capture human intuition. Kovatchev et al. [2018a] emphasize the similarity between linguistic phenomena underlying paraphrasing and entailment. There has been practical work on using paraphrasing to solve entailment [Bosma and Callison-Burch, 2006].

Interaction between Entailment and Specificity Specificity was involved in rules for the recognition of textual entailment [Bobrow et al., 2007].

Interaction with Semantic Similarity Cer et al. [2017] argue that to find paraphrases or entailment, some level of semantic similarity must be given. Furthermore, Cer et al. [2017] state that although semantic similarity includes both entailment and paraphrasing, it is different, as it has a gradation and not a binary measure of the semantic overlap. Based on their corpus, Marelli et al. [2014] state that paraphrases, entailment, and contradiction have a high similarity score; paraphrases having the highest and contradiction the lowest of them. There also was practical work using the interaction between semantic similarity and entailment: Yokote et al. [2011] and Castillo and Cardenas [2010] used semantic similarity to solve entailment.

7.2.2 Corpora with Multiple Semantic Layers

There are several works describing the creation, annotation, and subsequent analysis of corpora with multiple parallel phenomena.

MASC The annotation of corpora with multiple phenomena in parallel has been most notably explored within the Manually Annotated Sub-Corpus (MASC) project[2] — It is a large-scale, multi-genre corpus manually annotated with multiple semantic layers, including WordNet senses[Miller, 1998], Penn Treebank Syntax [Marcus et al., 1993], and opinions. The multiple layers enable analyses between several phenomena.

SICK is a corpus of around 10,000 sentence pairs that were annotated with semantic similarity and entailment in parallel [Marelli et al., 2014]. As it is the corpus that is the most similar to our work, we will compare some of our annotation decisions and results with theirs.

Sukhareva et al. [2016] annotated subclasses of entailment, including *paraphrase, forward, revert,* and *null* on propositions extracted from documents on educational topics that were paired according to semantic overlap. Hence, they implicitly regarded paraphrases as a kind of entailment.

[2] `http://www.anc.org/MASC/About.html`

7.3 Corpus Creation

To analyze the interactions between semantic relations, a corpus annotated with all relations in parallel is needed. Hence, we develop a new corpus-creation methodology which ensures all relations of interest to be present. First, we create a pool of potentially related sentences. Second, based on the pool of sentences, we create sentence pairs that contain all relations of interest with sufficient frequency. This contrasts existing corpora on meaning relations that are tailored towards one relation only. Finally, we take a portion of the corpus and annotate all relations via crowdsourcing. This part of our methodology differs significantly from the approach taken in the SICK corpus [Marelli et al., 2014]. They don't create new corpora, but rather re-annotate pre-existing corpora, which does not allow them to control for the overall similarity between the pairs.

7.3.1 Sentence Pool

Table 7.1 List of given source sentences

Getting a high educational degree is important for finding a good job, especially in big cities.
In many countries, girls are less likely to get a good school education.
Going to school socializes kids through constant interaction with others.
One important part of modern education is technology, if not the most important.
Modern assistants such Cortana, Alexa, or Siri make our everyday life easier by giving quicker access to information.
New technologies lead to asocial behavior by e.g. depriving us from face-to-face social interaction.
Being able to use modern technologies is obligatory for finding a good job.
Self-driving cars are safer than humans as they don't drink.
Machines are good in strategic games such as chess and Go.
Machines are good in communicating with people.
Learning a second language is beneficial in life.
Speaking more than one language helps in finding a good job.
Christian clergymen learn Latin to read the bible.

In the first step, the authors create 13 sentences, henceforth *source sentences*, shown in Table 7.1. The sentences are on three topics: *education*, *technology*, and *language*. We choose sentences that can be understood by a competent speaker

without any domain-specific knowledge and which due to their complexity potentially give rise to a variety of lexically differing sentences in the next step. Then, a group of 15 people, further on called *sentence generators*, is asked to generate *true* and *false* sentences that vary lexically from the source sentence.[3] Overall, 780 sentences are generated. The 13 *source sentences* are not considered in the further procedure.

For creating the *true* sentences, we ask each sentence generator to create two sentences that are true and for the *false* sentences, two sentences that are false given one source sentence. This way of generating a sentence pool is similar to that of the textual entailment SNLI corpus [Bowman et al., 2015], where the generators were asked to create true and false captions for given images. The following are exemplary true and false sentences created from one source sentence.

> **Source:** *Getting a high educational degree is important for finding a good job, especially in big cities.*
>
> **True:** *Good education helps to get a good job.*
>
> **False:** *There are no good or bad jobs.*

7.3.2 Pair Generation

We combine individual sentences from the sentence pool into pairs, as meaning relations are present between pairs and not individual sentences. To obtain a corpus that contains all discussed meaning relation with sufficient frequency, we use four pair combinations:

1) a pair of two sentences that are true given the same source sentence (*true-true*)

2) a pair of two sentences that are false given the same source sentence (*false-false*)

3) a pair of one sentence that is true and one sentence that is false given the same source sentence (*true-false*)

4) a pair of randomly matched sentences from the whole sentence pool and all source sentences (*random*)

From the 780 sentences in the sentence pool, we created a corpus of 11,310 pairs, with a pair distribution as follows: 5,655 (50%) *true-true*; 2,262 (20%)

[3]The full instructions given to the sentence generators is included with the corpus data.