# Annotating and analyzing the interactions between meaning relations

**Darina Gold[1*], Venelin Kovatchev[23*], Torsten Zesch[1]**
[1]Language Technology Lab, University of Duisburg-Essen, Germany
[2]Language and Computation Center, Universitat de Barcelona, Spain
[3]Institute of Complex Systems, Universitat de Barcelona, Spain
{darina.gold,torsten.zesch}@uni-due.de
vkovatchev@ub.edu
[*]Both authors contributed equally to this work

## Abstract

Pairs of sentences, phrases, or other text pieces can hold semantic relations such as paraphrasing, textual entailment, contradiction, specificity, and semantic similarity. These relations are usually studied in isolation and no dataset exists where they can be compared empirically. Here we present a corpus annotated with these relations and the analysis of these results. The corpus contains 520 sentence pairs, annotated with these relations. We measure the annotation reliability of each individual relation and we examine their interactions and correlations. Among the unexpected results revealed by our analysis is that the traditionally considered direct relationship between paraphrasing and bi-directional entailment does not hold in our data.

## 1 Introduction

Meaning relations refer to the way in which two sentences can be connected, e.g. if they express approximately the same content, they are considered paraphrases. Other meaning relations we focus on here are textual entailment and contradiction[1] (Dagan et al., 2005), and specificity.

Meaning relations have applications in many NLP tasks, e.g. recognition of textual entailment is used for summarization (Lloret et al., 2008) or machine translation evaluation (Padó et al., 2009), and paraphrase identification is used in summarization (Harabagiu and Lacatusu, 2010).

The complex nature of the meaning relations makes it difficult to come up with a precise and widely accepted definition for each of them. Also, there is a difference between theoretical definitions and definitions adopted in practical tasks. In this paper, we follow the approach taken in previous annotation tasks and we give the annotators generic and practically oriented instructions.

**Paraphrases** are differently worded texts with approximately the same content (Bhagat and Hovy, 2013; De Beaugrande and Dressler, 1981). The relation is symmetric. In the following example, (a) and (b) are paraphrases.

(a) *Education is equal for all children.*

(b) *All children get the same education.*

**Textual Entailment** is a directional relation between pieces of text in which the information of the *Text* entails the information of the *Hypothesis* (Dagan et al., 2005). In the following example, Text (t) entails Hypothesis (h):

(t) *All children get the same education.*

(h) *Education exists.*

**Specificity** is a relation between phrases in which one phrase is more precise and the other more vague. Specificity is mostly regarded between noun phrases (Cruse, 1977; Enç, 1991; Farkas, 2002). However, there has also been work on specificity on the sentence level (Louis and Nenkova, 2012). In the following example, (c) is more specific than (d) as it gives information on who does not get good education:

(c) *Girls do not get good education.*

(d) *Some children do not get good education.*

**Semantic Similarity** between texts is not a meaning relation in itself, but rather a gradation of meaning similarity. It has often been used as a proxy for the other relations in applications such as summarization (Lloret et al., 2008), plagiarism detection (Alzahrani and Salim, 2010; Bär et al., 2012), machine translation (Padó et al.,

[1]Mostly, contradiction is regarded as one of the relations within an entailment annotation.



*Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36
Florence, Italy, August 1, 2019. 

2009), question answering (Harabagiu and Hickl, 2006), and natural language generation (Agirre et al., 2013). We use it in this paper to quantify the strength of relationship on a continuous scale. Given two linguistic expressions, semantic text similarity measures the degree of semantic equivalence (Agirre et al., 2013). For example, (a) and (b) have a semantic similarity score of 5 (on a scale from 0-5 as used in the SemEval STS task) (Agirre et al., 2013, 2014).

**Interaction between Relations** Despite the interactions and close connection of these meaning relations, to our knowledge, there exists neither an empirical analysis of the connection between them nor a corpus enabling it. We bridge this gap by creating and analyzing a corpus of sentence pairs annotated with all discussed meaning relations.

Our analysis finds that previously made assumptions on some relations (e.g. paraphrasing being bi-directional entailment (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010; Sukhareva et al., 2016)) are not necessarily right in a practical setting. Furthermore, we explore the interactions of the meaning relation of specificity, which has not been extensively studied from an empirical point of view. We find that it can be found in pairs on all levels of semantic relatedness and does not correlate with entailment.

## 2 Related Work

To our knowledge, there is no other work where the discussed meaning relations have been annotated separately on the same data, enabling an unbiased analysis of the interactions between them. There are corpora annotated with multiple semantic phenomena, including meaning relations.

### 2.1 Interactions between relations

There has been some work on the interaction between some of the discussed meaning relations, especially on the relation between entailment and paraphrasing, and also on how semantic similarity is connected to the other relations.

**Interaction between entailment and paraphrases** According to Madnani and Dorr (2010); Androutsopoulos and Malakasiotis (2010), bi-directional entailment 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. (2018) 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.

### 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 doc-

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

| |
|---|
| 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. |

Table 1: List of given source sentences

uments on educational topics that were paired according to semantic overlap. Hence, they implicitly regarded paraphrases as a kind of entailment.

## 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.

### 3.1 Sentence Pool

In the first step, the authors create 13 sentences, henceforth *source sentences*, shown in Table 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.*

### 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*.

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

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%) *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.

### 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., 2005) - (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., 2005), 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

[4] Louis and Nenkova (2012) labelled individual sentences

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]

### 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 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 2 shows the overall inter-annotator agreement for the binary tasks. We report: 1) the average %-agreement for the whole corpus; 2) the average $\kappa$ 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 % agreement between the annotators and the expert-provided "control labels" on the control questions.

| | % | $\kappa$ | %✓ | %✗ | 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 2: Inter-annotator agreement for binary relations
✓denotes a relation being there
✗denotes a relation not being there

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.

| | 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 3: Distribution of Inter-annotator agreement

Table 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

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]We are aware that $\kappa$ 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.

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 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 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 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.

| | 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 |

Table 4: Distribution of meaning relations within different pair generation patterns

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 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 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.

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

Table 5: Comparison of BLEU scores between the sentence pairs in different corpora

**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 8 are examples for pairs containing negation in our corpus.

## 4 Interactions between relations

We analyze the interactions between the relations in our corpus in two ways. First, we calculate the

[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.

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.

### 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.

#### 4.1.1 Correlation of binary meaning relations

In Table 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 pairs that are both paraphrases and textual entailment are pairs #1 and #2 in Table 8. Furthermore, both paraphrases and entailment have a negative correlation with contradiction, which is expected and confirms the quality of our data.

Specificity does not have any strong correlation with any of the other relations, showing that it is independent of those in our corpus.

| | TE | UTE | BiTE | Cont | Spec | ∅ Sim |
|---|---|---|---|---|---|---|
| PP | .75 | .20 | .70 | -.25 | -.01 | 3.77 |
| TE | | .57 | .66 | -.30 | -.01 | 3.59 |
| UTE | | | -.23 | -.17 | -.04 | 3.21 |
| BiTE | | | | -.20 | -.01 | 3.89 |
| Cont | | | | | -.09 | 1.45 |
| Spec | | | | | | 2.27 |

Table 6: Correlation between all relations

#### 4.1.2 Binary relations and semantic similarity

We look at the average similarity for each relation (see Table 6) and show boxplots between relation labels and similarity ratings (see Figure 1). Table 6 shows that bi-directional entailment has the highest average similarity, followed by paraphrasing, while contradiction has the lowest.

Figure 1 shows plots of the semantic similarity for all pairs where each relation is present and all pairs where it is absent. The paraphrase pairs have much higher similarity scores than the non-paraphrase pairs. The same observation can be made for entailment. The contradiction pairs have a low similarity score, whereas the non-contradiction pairs do not have a clear tendency with respect to similarity score. In contrast to the other relations, pairs with and without specificity do not have any consistent similarity score.

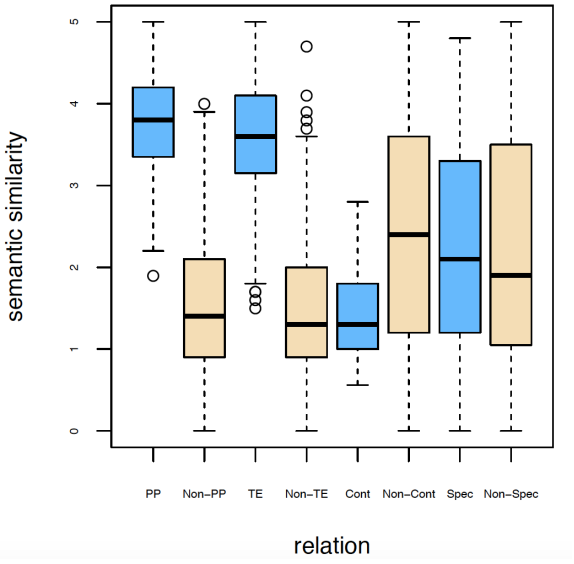


Figure 1: Similarity scores of sentences annotated with different relations

### 4.2 Overlap of relation labels

Table 7 shows the overlap between the different binary labels. Unlike Pearson correlation, the overlap is asymmetric - the % of paraphrases that are also entailment (UTE in PP) is different from the % of entailment pairs that are also paraphrases (PP in UTE). Using the overlap measure, we can identify interesting interactions between phenomena and take a closer look at some examples.

| | PP | UTE | BiTE | Contra | Spec |
|---|---|---|---|---|---|
| In PP | | 28 % | 64 % | 0 | 73 % |
| In UTE | 52 % | | - | 0 | 73 % |
| In BiTE | 94 % | - | | 0 | 72 % |
| In Contra | 0 | 0 | 0 | | 63 % |
| In Spec | 30 % | 17 % | 21 % | 11 % | |

Table 7: Distribution of overlap within relations

#### 4.2.1 Entailment and paraphrasing overlap

In a more theoretical setting, bi-directional entailment is often defined as being paraphrases (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010; Sukhareva et al., 2016). This

| # | Sentence 1 | Sentence 2 | PP | FTE | BTE | Cont | FSpec | BSpec | Sim |
|---|---|---|---|---|---|---|---|---|---|
| 1 | The importance of technology in modern education is overrated. | Technology is not mandatory to improve education | ✓ | ✓ | ✓ | | | | 2.8 |
| 2 | Machines cannot interact with humans. | No machine can communicate with a person. | ✓ | ✓ | ✓ | | | | 4.9 |
| 3 | The modern assistants make finding data slower. | Today's information flow is greatly facilitated by digital assistants. | | | | ✓ | | ✓ | 1.9 |
| 4 | The bible is in Hebrew. | Bible is not in Latin. | | ✓ | | | ✓ | | 2.7 |
| 5 | All around the world, girls have higher chance of getting a good school education. | Girls get a good school education everywhere. | ✓ | | | | | ✓ | 4.7 |
| 6 | Reading the Bible requires studying Latin. | The Bible is written in Latin. | | ✓ | ✓ | | | ✓ | 3.6 |
| 7 | Speaking more than one language can be useful. | Languages are beneficial in life. | ✓ | ✓ | ✓ | | | ✓ | 4.4 |
| 8 | You can find a good job if you only speak one language. | People who speak more than one language could only land pretty bad jobs. | | | ✓ | | | | 2.3 |
| 9 | All Christian priests need to study Persian, as the Bible is written in Ancient Greek. | Christian clergymen don't read the bible. | | | | | | ✓ | 0.9 |
| 10 | School makes students antisocial. | School usually prevents children from socializing properly. | ✓ | ✓ | ✓ | | | ✓ | 3.9 |

Table 8: Annotations of sentence pairs on all meaning relations taken from our corpus

implies that paraphrases equal bi-directional entailment. In our corpus, we can see that only 64% of the paraphrases are also annotated as bi-directional entailment. An example of a pair that is annotated both as paraphrase and as bi-directional entailment is pair #10 in Table 8. However, in the corpus we also found that 28 % of the paraphrases are only uni-directional entailment, while in 8% annotators did not find any entailment. An example of a pair where our annotators found paraphrasing, but not entailment is sentence pair #5 in Table 8. The agreement on the paraphrasing for this pair was 80%, the agreement on (lack of) forward and backward entailment was 80% and 70% respectively. Although the information in both sentences is nearly identical, there is no entailment, as "having a higher chance of getting smth" does not entail "getting smth" and vice versa.

If we look at the opposite direction of the overlap, we can see that 52% of the uni-directional and 94% of the bi-directional entailment pairs are also paraphrases. This finding confirms the statement that bi-directional entailment is paraphrasing (but not vice versa).

There is also a small portion (6%) of bi-directional entailments that were not annotated as paraphrases. An example of this is pair #6 in Table 8. Although both sentences make each other true, they do not have the same content.

Neither paraphrasing nor entailment had any overlap with contradiction, which further verifies our annotation scheme and quality.

These findings are partly due to the more "relaxed" definition of paraphrasing adopted here. Our definition is consistent with other authors that work on paraphrasing and the task of paraphrase identification, so we argue that our findings are valid with respect to the practical applications of paraphrasing and entailment and their interactions.

#### 4.2.2 Overlap with specificity

Specificity has a nearly equal overlap within all the other relations. In the pairs annotated with paraphrase or entailment, 73% are also annotated with specificity. The high number of pairs that are in a paraphrase relation, but also have a difference in specificity is interesting, as it seems more natural for paraphrases to be on the same specificity level. One example of this is pair #7 in Table 8. Although they are paraphrases (with 100% agreement), the first one is more specific, as it 1) specifies the ability of speaking a language and 2) says "more than one language".

There are also 27% of uni-directional entailment relation pairs that are not in any specificity relation. One example of this is pair #8 in Table 8.

Although the pair contains uni-directional entailment (backward entailment), none of the sentences is more specific than the other.

If we look at the other direction of the overlap, we can observe that in 62% of the cases involving difference in specificity, there is no uni-directional nor bi-directional entailment. An example of such a relation pair is pair #9 in Table 8. The two sentences are on the same topic and thus can be compared on their specificity. The first sentence is clearly more specific, as it gives information on what needs to be learned and where the Bible was written, whereas the second one just gives an information on what Christian clergymen do. These findings indicate that entailment is not specificity.

### 4.3 Discussion

Our methodology for generating text pairs has proven successful in creating a corpus that contains all relations of interest. By selecting different sentence pairings, we have obtained a balance between the relations that best suit our needs.

The inter-annotator agreement was good for all relations. The resulting corpus can be used to study individual relations and their interactions. It should be emphasized that our findings strongly depend on our decisions concerning the annotations setup, the guidelines in particular. When examining the interactions between the different relations, we found several interesting tendencies.

**Findings on the interaction between entailment and paraphrases** We showed that paraphrases and any-directional entailment had a high correlation, high overlap, and a similarly high semantic similarity. Almost all bi-directional entailment pairs are paraphrases. However, only 64% of the paraphrases are bi-directional entailment, indicating that paraphrasing is the more general phenomena, at least in practical tasks.

**Findings on specificity** With respect to specificity, we found that it does not correlate with other relations, showing that it is independent of those in our corpus. It also shows no clear trend on the similarity scale and no correlation with the difference in word length between the sentences. This indicates that specificity cannot be automatically predicted using the other meaning relations and requires further study.

In the examples that we discuss, we focus on interesting cases, which are complicated and unexpected (ex.: paraphrases that are not entailment or entailment pairs that do not differ in specificity). However, the full corpus also contains many conventional and non-controversial examples.

## 5 Conclusion and Further Work

In this paper, we made an empirical, corpus-based study on interactions between various semantic relations. We provided empirical evidence that supports or rejects previously hypothesized connections in practical settings. We release a new corpus that contains all relations of interest and the corpus creation methodology to the community. The corpus can be used to further study relation interactions or as a more challenging dataset for detecting the different relations automatically[10].

Some of our most important findings are:

1) there is a strong correlation between paraphrasing and entailment and most paraphrases include at least uni-directional entailment;

2) paraphrases and bi-directional entailment are not equivalent in practical settings;

3) specificity relation does not correlate strongly with the other relations and requires further study;

4) contradictions (in our dataset) are perceived as dis-similar.

As a future work, we plan to: 1) study the specificity relation in a different setting; 2) use a linguistic annotation to determine more fine-grained distinctions between the relations; 3) and annotate the rest of the 11,000 sentences in a semi-automated way.

## Acknowledgements

We would like to thank Tobias Horsmann and Michael Wojatzki and the anonymous reviewers for their suggestions and comments. Furthermore, we would like to thank the sentence generators for their time and creativity. This work has been partially funded by Deutsche Forschungsgemeinschaft within the project ASSURE. This work has been partially funded by Spanish Ministery of Economy Project TIN2015-71147-C2-2, by the CLiC research group (2017 SGR 341), and by the APIF grant of the second author.

[10]The full corpus, the annotation guidelines, and the control examples can be found at https://github.com/MeDarina/meaning_relations_interaction

## References


Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. pages 81–91.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. volume 1, pages 32–43.

Salha Alzahrani and Naomie Salim. 2010. Fuzzy semantic-based string similarity for extrinsic plagiarism detection. *Braschler and Harman* 1176:1–8.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research* 38:135–187.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2012. Text reuse detection using a composition of text similarity measures. *Proceedings of COLING 2012* pages 167–184.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics* 39(3):463–472.

Daniel Bobrow, Dick Crouch, Tracy Halloway King, Cleo Condoravdi, Lauri Karttunen, Rowan Nairn, Valeria de Paiva, and Annie Zaenen. 2007. Precision-focused textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. pages 16–21.

Wauter Bosma and Chris Callison-Burch. 2006. Paraphrase substitution for recognizing textual entailment. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, pages 502–509.

Samuel R Bowman, 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.

Julio J. Castillo and Marina E. Cardenas. 2010. Using sentence semantic similarity based on WordNet in recognizing textual entailment. In *Ibero-American Conference on Artificial Intelligence*. Springer, pages 366–375.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. pages 1–14.

D. Alan Cruse. 1977. The pragmatics of lexical specificity. *Journal of linguistics* 13(2):153–164.

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

Robert De Beaugrande and Wolfgang U Dressler. 1981. *Introduction to text linguistics*. Routledge.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 350.

Mürvet Enç. 1991. The semantics of specificity. *Linguistic inquiry* pages 1–25.

Donka F. Farkas. 2002. Specificity distinctions. *Journal of semantics* 19(3):213–243.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 905–912.

Sanda Harabagiu and Finley Lacatusu. 2010. Using topic themes for multi-document summarization. *ACM Transactions on Information Systems (TOIS)* 28(3):13.

Cordeiro Joao, Dias Gaël, and Brazdil Pavel. 2007. New functions for unsupervised asymmetrical paraphrase detection. *Journal of Software* 2(4):12–23.

Venelin Kovatchev, M. Antónia Martì, and Maria Salamo. 2018. ETPC - A Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. 2008. A Text Summarization Approach under the Influence of Textual Entailment. In *NLPCS*. pages 22–31.

Annie Louis and Ani Nenkova. 2012. A corpus of general and specific sentences from news. In *LREC*. pages 1818–1821.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics* 36(3):341–387.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank .

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*. pages 216–223.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D Manning. 2009. Textual entailment features for machine translation evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 37–41.

Maria Sukhareva, Judith Eckle-Kohler, Ivan Habernal, and Iryna Gurevych. 2016. Crowdsourcing a Large Dataset of Domain-Specific Context-Sensitive Semantic Verb Relations. In *LREC*.

Ken-ichi Yokote, Shohei Tanaka, and Mitsuru Ishizuka. 2011. Effects of Using Simple Semantic Similarity on Textual Entailment Recognition. In *TAC*.