We address these research questions in a corpus annotation study. For the first research question we evaluate the quality of the corpus annotation by measuring the inter-annotator agreement. For the second research question we measure the relative frequencies of the types in sentence pairs with each textual meaning relation.

8.4 Corpus Annotation

This section is organized as follows: Section 8.4.1 describes the corpus that we chose to use in the annotation. Section 8.4.2 presents the annotation setup. Finally, in Section 8.4.3 we report the annotation agreement.

8.4.1 Choice of Corpus

In order to determine the applicability of SHARel to all relations of interest, we carried out a corpus annotation. We used the publicly available corpus of Gold et al. [2019]. It consists of 520 text pairs and is already annotated at sentence level for paraphrasing, entailment, contradiction, specificity and semantic similarity. Gold et al. [2019] performed the annotation for each relation independently. That is, for each pair of sentences 10 annotators were asked whether a particular relation (paraphrasing, entailment, contradiction, specificity) held or not.

The corpus of Gold et al. [2019] contains 160 pairs annotated as paraphrases, 195 pairs annotated as textual entailment (in one direction or in both) and 68 pairs annotated as contradiction. As the annotation for the different relations was carried out independently, there is an overlap between the relations. For example 52% of the pairs annotated as entailment were also annotated as paraphrases. The total number of pairs annotated with at least one relation among paraphrasing, entailment, and contradiction is 256. The remaining 244 pairs were annotated as unrelated. In 381 of the pairs, one of the sentences was marked as more specific than the other.

The corpus of Gold et al. [2019] is the only corpus to date which contains all relations of interest. All text pairs are in the same domain and topic, they have similar syntactic structure and vocabulary. The lexical overlap between the two sentences in each pair is much lower than in corpora such as MRPC [Dolan et al., 2004] or SNLI [Bowman et al., 2015]. This means that even though the two sentences in a pair are in a meaning relation such as paraphrasing or textual entailment, there are very few words that are directly repeated. All these properties of the corpus were taken into consideration when we chose it for our annotation.

8.4.2 Annotation Setup

We performed an annotation with the SHARel typology on all pairs from Gold et al. [2019] that have at least one of the following relations: paraphrasing, forward entailment, backwards entailment, and contradiction. We discarded pairs that are annotated as "unrelated". This is a typical approach when decomposing meaning relations. Sammons et al. [2010], Cabrio and Magnini [2014], Vila et al. [2014] only decompose pairs with a particular relation (entailment, contradiction, or paraphrasing).

After discarding the unrelated portion, the total number of pairs that we annotated with SHARel was 276. Prior to the annotation we tokenized each sentence using the NLTK python library.

During the annotation process, our annotators go through each pair in the corpus. For each linguistic and reason-based phenomenon that they encounter, they annotate the type and the scope (the specific tokens affected by the type). We used an open source web-based annotation interface, called WARP-Text [Kovatchev et al., 2018b].

We prepared extended guidelines with examples for each type. Each pair of texts was annotated independently by two trained expert annotators. In the cases where there were disagreements, the annotators discussed their differences in order to obtain the best possible annotation for the example pair [3].

8.4.3 Agreement

For calculating inter-annotator agreement, we use the two different versions of the IAPTA-TPO measures. The IAPTA-TPO measures was proposed by Vila et al. [2015] specifically for the task of annotating paraphrase types. They were later on refined by Kovatchev et al. [2018a]. IAPTA-TPO measure the agreement on both the label (the annotated phenomenon) and the scope, which is non-trivial to capture using traditional measures such as Kappa. IAPTA-TPO (Total) measures the cases where the annotators fully agree on both label and scope. IAPTA-TPO (Partial) measures the cases where the annotators agree on the label, but the scope overlaps only partially.

The agreement of our annotation can be seen in Table 8.3. We calculate the agreement on all pairs (all), and we also report the agreement for the pairs with textual label paraphrases (pp), entailment (ent), and contradiction (cnt).

[3]The annotation guidelines and the annotated corpus are available at `https://github.com/venelink/sharel`

Table 8.3 Inter-annotator Agreement

	TPO-Partial	**TPO-Total**
This corpus (all)	.78	.52
This corpus (pp)	.77	.51
This corpus (ent)	.77	.52
This corpus (cnt)	.75	.50
MRPC-A	.78	.51
ETPC (non-pp)	.72	.68
ETPC (pp)	.86	.68

To put our results in perspective, we compare our agreement with the one reported in MRPC-A [Vila et al., 2015] and ETPC [Kovatchev et al., 2018a]. For ETPC the authors report both the agreement on the pairs annotated as paraphrases (pp) and as non-paraphrases (non-pp). To date, MRPC-A and ETPC are the only two corpora of sufficient size annotated with a typology of meaning relations. They also use the same inter-annotation measure to report agreement, so we can compare with them directly.

The overall agreement that we obtain (.52 Total and .78 Partial) is almost identical to the agreement reported for MRPC-A (.51 Total and .78 Partial) and slightly lower than the agreement reported for ETPC (.68 Total and .86 Partial).

Kovatchev et al. [2018a] detected a significant difference in the agreement between paraphrase and non-paraphrase pairs. In their annotation, the "non-paraphrase" includes mostly entailment and contradiction pairs and the lower agreement indicates that their typology is not well equipped for dealing with those cases. However in our corpus, we don't observe such a difference. Our annotation agreement is very consistent across all pairs indicating that SHARel is successfully applied to all relations of interest.

The consistently high agreement score indicates the high quality of the annotation. Even though our task and our typology are much more complex than those of Vila et al. [2014] and Kovatchev et al. [2018a], we still obtain comparable results.

In addition to calculating the inter-annotation agreement, we also asked the annotators to mark and indicate any examples and/or phenomena not covered by the typology. Based on their ongoing feedback during the annotation, we decided to introduce the "anaphora" type. We re-annotated the portion of the corpus that was already annotated at the time when we introduced the new type.

Arriving at this point, we have demonstrated that it is possible to successfully use a single typology for the decomposition of multiple (textual) meaning relations. This answers our first research question (**RQ1**).