Table 5.2 Non-sense preserving phenomena

type	pair	s1 scope	s2 scope	s1 text	s2 text
7	146	11	11	Friday	Thursday
10	146	7	8	slip	rise

The annotation of 12a and 12b illustrates one of the issues when annotating non-sense preserving phenomena. In many textual pairs, there is more than one "key" difference. In those cases, all of the phenomena were annotated separately. Nevertheless, the annotators were instructed to be conservative and only annotate phenomena that carry substantial differences in the meaning of the two texts. Determining which differences are substantial, and which are not was the main challenge for the annotators. Due to the difficulty of the task, we selected annotators that were expert linguists with a high proficiency of English[6].

When the two texts were substantially different and it was not possible to identify the atomic phenomena responsible for the difference, the pair was annotated with atomic type *"non-paraphrase"* (examples 13a and 13b) or *"entailment"* (examples 14a and 14b).

13a That compared with $35.18 million, or 24 cents per share, in the year-ago period.

13b Earnings were affected by a non-recurring $8 million tax benefit in the year-ago period.

14a The year-ago comparisons were restated to include Compaq results.

14b The year-ago numbers do not include figures from Compaq Computer.

5.4.2 Sense Preserving Atomic Phenomena

For the annotation of the sense preserving atomic phenomena, we used the same annotation scheme format as the one for the non-sense preserving phenomena. Each phenomenon is identified by a type, a scope, and, where applicable, a key. 15a and 15b show a textual pair, annotated as a paraphrase in the MRPC. An example of a single annotated atomic phenomenon can be seen in Table 5.3

15a Amrozi accused his brother , whom he called " the witness " , of deliberately distorting his evidence .

[6]The full annotation guidelines for both sense preserving and non-sense preserving phenomena can be found at `https://github.com/venelink/ETPC` and in Appendix A of the thesis.

15b Referring to him as only “ the witness ” , Amrozi accused his brother of deliberately distorting his evidence .

Table 5.3 Sense preserving phenomenon

type	pair	s1 scope	s2 scope	s1 text	s2 text
6	1	5	1, 2	whom	to him

For the 3900 text pairs already annotated by Vila et al. [2015], we worked with the existing corpus and we only re-annotated the 3 new sense preserving paraphrase types introduced in EPT. For the 1901 textual non-paraphrases, which were not annotated in MRPC-A, we performed a full annotation with all 25 sense preserving atomic types.

5.4.3 Inter-Annotator Agreement

In this section, we present the measures for calculating the inter-annotator agreement and the agreement score on the first two layers of annotation: non-sense preserving atomic phenomena and sense preserving atomic phenomena.

The measure that we use is the IAPTA TPO, introduced by Vila et al. [2015]. It is a fine-grained measure, created specifically for the task of annotating paraphrase types. It takes into account the agreement with respect to both the label and the scope of the phenomena. It is a pairwise agreement measure, obtained by calculating the Precision, Recall and F1 of one of the annotators, while using another annotator as a gold standard. There are two versions of the measure - TPO-partial, which requires that the annotators select the same label and that the scopes overlap by at least one token; and TPO-total which requires full overlap of label and scope.

The classical TPO measures are pairwise, they calculate the agreement between two annotators. When the annotation process involves more than two annotators, we first calculate the pairwise TPO measure between any two annotators and then we use one of three different techniques for calculating the overall agreement for the corpus. TPO (avg) is the most simple score, as it is the average of all pairwise TPO scores. TPO (union) is the union of all pairwise TPO agreement tables. That is, any phenomena that is annotated with the same label and the same scope by any 2 annotators is part of the TPO (union). Finally, TPO (gold) is the average F1 score of the three annotators, when treating TPO (union) as a gold standard. TPO (union) and TPO (gold) are two new measures, that we propose as part of this paper. TPO (union) represents all the “high quality” phenomena (that

is, phenomena annotated the same way by multiple annotators). TPO (gold) represents the probability that any of our annotators would annotate "high quality" phenomena.

The annotation of the sense preserving atomic paraphrases was carried out by two expert annotators, while the annotation of the non-sense preserving atomic phenomena was carried out by three expert annotators. For the purpose of calculating the inter-annotator agreement, all experts were given the same 180 text pairs (roughly 10 % of all non-paraphrase pairs in the corpus). The pairs were split in 3 equal parts and given to the annotators in three different stages of the annotation: one at the beginning, one in the middle, and one at the end of the annotation process. Table 5.4 shows the obtained scores, where ETPC (-) stands for the non-sense preserving layer, ETPC (+) stands for the sense-preserving layer of annotation and MRPC-A is the annotation of Vila et al. [2015]. For ETPC (+) we only had two annotators, so we were not able to calculate TPO (union) and TPO (gold). Since these measures have been introduced by us in the current paper, the MRPC-A corpus by Vila et al. [2015] does not have values for them either.

Table 5.4 Inter-annotator Agreement

Measure	ETPC (-)	ETPC (+)	MRPC-A
TPO-partial (avg)	0.72	0.86	0.78
TPO-total (avg)	0.68	0.68	0.51
TPO (union)	0.77	n-a	n-a
TPO (gold)	0.86	n-a	n-a

ETPC (+) and MRPC-A are directly comparable as they measure the agreement on the same task (annotation of sense-preserving atomic phenomena). The results show much higher agreement score with respect to both TPO-partial (0.86 against 0.78) and TPO-total (0.68 against 0.51). ETPC (-) measures the agreement on a different task (annotation of non-sense preserving phenomena). The TPO-partial score of ETPC (-) is lower than both ETPC (+) and MRPC-A (0.72 against 0.86 and 0.78 respectively), however the TPO-total score is equal to that of ETPC (+) and much higher than that of MRPC-A. It is interesting to note that there is almost no difference between TPO-partial and TPO-total for ETPC (-) (0.72 against 0.68), while for ETPC (+) and MRPC-A, the difference is significant. The TPO (union) for ETPC (-) shows that 77% of all phenomena are annotated the same way by at least 2 of the annotators. The TPO (gold) indicates that the probability of any of our experts annotating a "gold" example is 86%. Considering the difficulty of the task, the obtained results indicate the high quality of the annotated corpus.