5.4.4 Annotation of Negation

During the first two steps of the annotation, we identified all sentences that contain negation. For every instance of negation we annotated the negation cues and the scope of negation. 16a and 16b illustrate an example of annotated negation.

16a (Moore had (**no** [negation marker]) immediate comment Tuesday [scope])

16b (Moore (**did not** [negation marker]) have an immediate response Tuesday [scope])

5.5 The ETPC Corpus

This section presents the results of the annotation of the ETPC corpus. Section 5.5.1 shows the results of annotating non-sense preserving phenomena. Section 5.5.2 shows the results of annotating sense preserving phenomena. Section 5.5.3 discusses the results and the Research Questions, and Section 5.5.4 lists some applications of ETPC.

5.5.1 Non-Sense Preserving Atomic Phenomena

Table 5.5 shows the distribution of the non-sense preserving phenomena. Type Relative Frequency (Type RF) shows the relative distribution of the atomic types. Occurrence Frequency (Type OF) shows the distribution of phenomena per sentence, that is in how many textual pairs each phenomenon can be found[7]. The total number of non-sense preserving phenomena is 3406 in 1901 text pairs.

Both Type Relative Frequency (RF) and Occurrence Frequency (OF) indicate that the non-paraphrase portion of the corpus is not well balanced with respect to atomic phenomena. In 260 of the text pairs (13.7%), the annotators selected *“non-paraphrase”* indicating that the two texts were substantially different. In the rest of the pairs, the most common reason for the “non-paraphrase” label at textual level was *“Addition/Deletion”* (52% RF, 65.5% OF), followed by *“Same polarity substitution (named entity)”* (27% RF, 22.5% OF), *“Same polarity substitution (contextual)”* (RF 9,3%, OF 15.5%), and *“Opposite polarity substitution (habitual)”* (RF 2.8%, OF 4.6%). These are the only types with Type Relative Frequency and Occurrence Frequency above 1%, and they constitute over 99% of all non-sense preserving atomic phenomena annotated in the corpus. Six of the atomic phenomena are represented only with a few examples, while two are not represented at all.

[7] The sum of all Occurrence Frequencies exceeds 100, as one sentence often contains more than one atomic phenomenon.

Table 5.5 Distribution of non-sense preserving phenomena

Type	Type RF	Type OF
Inflectional	0.02%	0.04%
Same Polarity (con)	9.3%	15.5%
Same Polarity (ne)	27.5%	22.5%
Opp Polarity (hab)	2.7%	4.4%
Opp Polarity (con)	0.01%	0.02%
Converse	0.01%	0.02%
Diathesis	0.01%	0.01%
Negation	0.02%	0.03%
Direct/Indirect	0%	0%
Addition/Deletion	52%	65.5%
Semantic based	0%	0%
Non-paraphrase	7.6%	13.7%
Entailment	0.02%	0.04%

5.5.2 Sense Preserving Atomic Phenomena

Table 5.6 shows the distribution of sense preserving atomic phenomena in the textual paraphrase and non-paraphrase portions of the corpus[8]. For the textual paraphrase portion, we used the numbers reported by Vila et al. [2015] with partial re-annotation to account for the new types in ETPC. For *"same polarity substitution"*, 35% of the phenomena were re-annotated as *"habitual"*, 47% as *"contextual"*, and 18% as *"named entity"*. For *"opposite polarity substitution"* 21% of the phenomena were *"contextual"* and 79% of the phenomena were *"habitual"*.

The results show that the distribution of sense-preserving phenomena is relatively consistent between the two portions of the corpus. The most notable differences between the two distributions are the frequencies of *"same polarity substitution (named entity)"*, *"synthetic/analytic"*, *"addition/deletion"*, and *"identity"*. Both distributions are not well balanced in terms of atomic types, with 8 types (*"addition/deletion"*, *"identity"*, *"same polarity substitution (contextual)"*, *"same polarity substitution (habitual)"*, *"synthetic/analytic"*, *"same polarity substitution (named entity)"*, *"change of order"*, and *"punctuation"*) responsible for over 80% of the phenomena.

[8] At the time of the submission of this paper, the annotation of the non-paraphrase portion was not finished. The reported results are for 500 annotated pairs (about 30% of the corpus). The full figures will be made available at `https://github.com/venelink/ETPC`

Table 5.6 Distribution of Sense preserving phenomena in textual paraphrases and textual non-paraphrases

Type	Non Paraphrase	Paraphrase
Inflectional	2.13%	2.78%
Modal verb	0.59%	0.83%
Derivational	0.35%	0.85%
Spelling changes	1.30%	2.91%
Same Polarity (hab)	10.55%	8.68%
Same Polarity (con)	11.15%	11.66%
Same Polarity (ne)	7.11%	5.08%
Format	1.06%	1.1%
Opp Polarity (hab)	0%	0.07%
Opp Polarity (con)	0%	0.02%
Synthetic/analytic	7.82%	3.80%
Converse	0.12%	0.20%
Diathesis	0.83%	0.73%
Negation	0%	0.09%
Ellipsis	0.47%	0.30%
Coordination	0.24%	0.22%
Subord. and nesting	1.18%	2.14%
Punctuation	2.72%	3.77%
Direct/Indirect	0.24%	0.30%
Sentence modality	0%	0%
Synt./Disc. structure	1.30%	1.39%
Addition/Deletion	20.04%	25.94%
Change of order	3.08%	3.89%
Semantic	0%	1.53%
Identity	25.02%	17.54%
Non-Paraphrase	2.49%	3.81%
Entailment	0.12%	0.37%

5.5.3 Discussion

In this section we briefly discuss the annotation results and the Research Questions that we posed in Section 5.3.3

With respect to **RQ1** and **RQ2**, we measured the raw frequency distribution of the sense preserving atomic phenomena in both the paraphrase and non-