paraphrase portions of the corpus. We make two important observations from the data. First, the corpus is not well balanced in terms of type distribution in either of the portions. It can be seen in Table 5.6 that 8 of the types are overrepresented while the rest are underrepresented. This imbalance is even more significant in terms of meta-categories. The structure meta-types "syntax" and "discourse" account for less than 10 % of all types. Second, the raw frequency distribution of atomic phenomena in textual paraphrases and textual non-paraphrases is very similar. This finding suggests that it is the non-sense preserving phenomena that are mostly responsible for the relation at textual level in this corpus. This makes the annotation of the non-sense preserving phenomena even more important for the PI task.

With respect to **RQ3**, we annotated the *"same polarity substitution (contextual)"* and *"opposite polarity substitution (contextual) "* types in all portions of the corpus. For *"same polarity substitution"*, over 40% of the sense-preserving and over 25% of the non-sense preserving instances were contextual. For *"opposite polarity substitution"*, 21% of the sense-preserving instances were annotated as contextual, while in the non-sense preserving portion we found almost no contextual instances.

With respect to **RQ4**, we measured the raw frequency distribution of the non-sense preserving phenomena. If we compare it with the distribution of sense preserving phenomena, we can see that the differences are noteworthy and we can easily differentiate between the two distributions. Non-sense preserving phenomena are even less balanced than sense preserving phenomena, with just 4 types responsible for almost all instances. The structure types "syntax" and "discourse" are not represented at all, with all frequent types being either "lexical", "lexico-syntactic", or "other".

Finally, it is worht mentioning that 13% of the sentences in the textual paraphrase portion of the corpus and 12% of the sentences in the textual non-paraphrase portion contain negation. The relative distribution in the paraphrase and in the non-paraphrase portion of the corpus is consistent. The negation scope for each of these sentences has been annotated in a separate layer.

5.5.4 Applications of ETPC

The ETPC corpus has clear advantages over the currently available PI corpora, and the MRPC in particular. It is much more informative and can be used in several ways.

First, ETPC can be used as a single PI corpus. The annotation with atomic types makes it much more informative for evaluation than any other existing PI corpus. PI systems are currently evaluated in terms of binary Precision, Recall, F1 and Accuracy. ETPC provides the developers with much more detailed infor-

mation, without requiring any additional work on the developers' side. Knowing which atomic types are involved in the correct and incorrect classification helps the error analysis and should lead to an improvement in the these systems' performance. It also promotes reusability.

Second, ETPC can be used to provide quantitative and qualitative analysis of the MRPC corpus, as we have already shown in section 5.5.3 By having a detailed statistical analysis of the content of the corpus we can identify possible biases and promote the creation of better and more balanced corpora.

Third, ETPC can be easily split into various smaller corpora built around a certain atomic type or a class of types. Each of them can be used for a new task of Atomic Paraphrase Identification. It can be used to study the nature of the relation between atomic paraphrases and textual paraphrases.

Finally, ETPC can be used to study the role of negation in PI, a research question that, to date, has received very little attention.

5.6 Conclusions and Future Work

In this paper we presented the ETPC corpus - the largest corpus annotated with detailed paraphrase typology to date. For the annotation we used the new Extended Paraphrase Typology, a practically oriented typology of atomic paraphrases. The annotation process included three expert linguists and covered the whole 5801 text pairs from the MRPC corpus. The full corpus is publicly available in two formats: SQL and XML[9].

ETPC is a high quality resource for paraphrase related research and the task of PI. It provides more in-depth analysis of the existing corpora and promotes better understanding of the phenomena, the data, and the task. It also identifies several problems, such as the under-representation of structure based types and the over-representation of lexical based types. ETPC sets an example for the development of new feature-rich corpora for paraphrasing research. It also promotes collaboration between similar areas, such as PI, RTE and Semantic Similarity.

Our work opens several lines of future research. First, the ETPC can be used to re-evaluate existing state-of-the-art PI systems. This detailed evaluation can lead to improvements of the existing PI systems and the creation of new ones. Second, it can be used to create new corpora for paraphrase research, which will be more balanced in terms of type distribution. Third, it can be used to study the nature of the paraphrase phenomenon and the relation between "atomic" and "tex-

[9]We have also made publicly available all complementary data, such as annotation guidelines, screenshots of the interface, detailed statistics, as well as the ETPC_Neg corpus, composed only from the paraphrase and non-paraphrase pairs containing negation (`https://github.com/venelink/ETPC`).

tual" paraphrases. Finally, the EPT and ETPC can be extended to other research areas, such as lexical and textual entailment, semantic similarity, simplification, summarization, and question answering, among others.