whether a certain type can give raise to textual paraphrases (+), to textual non-paraphrases (-), or to both (+ / -)[4]. The typology contains 25 atomic paraphrase types (+) and 13 atomic non-paraphrase types (-). It is based on the work of Vila et al. [2014] and aims to extend it in two directions in order to address the four Research Questions.

First, we have added three new atomic paraphrase types - we split the atomic types *"same polarity substitution"* and *"opposite polarity substitution"* into two separate types based on the nature of the relation between the substituted words: *"habitual"* and *"contextual"*. We have also added the type *"same polarity substitution (named entity)"*. While the principle behind all substitutions is the same, in practice there is a significant difference whether the replaced words are connected in their habitual meaning, contextually, or refer to related named entities in the world. Instances of the new types can be seen in sentence pairs 5 (*"same polarity substitution (habitual)"*), 6 (*"same polarity substitution (contextual)"*), 7 (*"same polarity substitution (named entity"*), 8 (*"opposite polarity substitution (habitual)"*), and 9 (*"opposite polarity substitution (contextual)"*)

5a A federal magistrate in Fort Lauderdale ordered him held without bail.

5b Zuccarini was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla.

6a Meanwhile, the global death toll approached 770 with more than 8,300 people sickened since the severe acute respiratory syndrome virus first appeared in southern China in November.

6b The global death toll from SARS was at least 767, with more than 8,300 people sickened since the virus first appeared in southern China in November.

7a He told The Sun newspaper that Mr. Hussein's daughters had British schools and hospitals in mind when they decided to ask for asylum.

7b "Saddam's daughters had British schools and hospitals in mind when they decided to ask for asylum – especially the schools," he told The Sun.

8a Leicester failed in both enterprises.

8b He did not succeed in either case.

9a A big surge in consumer confidence has provided the only positive economic news in recent weeks.

[4] A more detailed table of EPT, with additional examples for each atomic type is available at `https://github.com/venelink/ETPC` and in Appendix A of the thesis.

9b Only a big surge in consumer confidence has interrupted the bleak economic news.

Second, we have introduced the *"sense preserving"* feature in 13 of the atomic types. As we have shown in the previous section (examples 4a and 4b), the same atomic linguistic transformation (such as substitution, diathesis alternation, and negation switching) can give raise to different semantic relations at textual level: paraphrasing, entailment, and contradiction, among others. This idea has already been expressed by Cabrio and Magnini [2014] in the field of Recognizing Textual Entailment. Building on this idea, we identify 13 atomic types that can, in different instances, give rise to both paraphrases and non-paraphrases. Sentence pairs 10 and 11 show an example of sense preserving and non-sense preserving *"Inflection change"* types. In 10a and 10b, both *"streets"* and *"street"* are a generalization with the meaning *"all streets"*. In a similar way, in 11b, *"boats"* has the meaning as *"all boats"*. However in 11a, *"boat"* can have the meaning *"one particular boat"*, thus the inflectional change *"boat - boats"* is not sense-preserving.

10a It was with difficulty that the course of streets could be followed.

10b You couldn't even follow the path of the street.

11a You can't travel from Barcelona to Mallorca with the boat.

11b Boats can't travel from Barcelona to Mallorca.

The changes introduced in EPT allow us to work on all four Research Questions (RQs) defined in Section 5.3.3 This is a clear advantage over the existing paraphrase typologies, which are only suitable for addressing **RQ1**. For **RQ1**, we annotated all atomic types in the positive ("paraphrases") portion of MRPC and measured their distribution. For **RQ2**, we annotated all atomic types in the negative ("non-paraphrases") portion of MRPC and compared the distribution of the types in the positive and negative portions. For **RQ3**, the two newly added "contextual" types allow us to distinguish and compare context dependent from context independent atomic paraphrases. Finally, for **RQ4**, the addition of "sense preserving" allows us to annotate, isolate and compare the sense preserving and non-sense preserving instances of the same linguistic phenomena.

5.4 Annotation Scheme and Guidelines

We propose the Extended Paraphrase Typology (EPT) with a clear practical objective in mind: to create language resources that improve the performance, evaluation, and understanding of the systems competing on the task of PI and to open new research directions. We used the EPT to annotate the MRPC corpus with atomic paraphrases. We annotated all 5801 text pairs in the corpus, including both the pairs annotated as paraphrases (3900 pairs) and those annotated as non-paraphrases (1901 pairs).

As a basis, we used the MRPC-A corpus by Vila et al. [2015], which already contains some annotated atomic paraphrases. Our annotation consisted of three steps, corresponding to the three different layers of annotation.

First, we annotated the non-sense preserving atomic phenomena (Section 5.4.1) in the textual non-paraphrases. Second, we annotated the sense preserving atomic paraphrase phenomena (Section 5.4.2) in both textual paraphrases and textual non-paraphrases. And third, we identified all sentences in the corpus containing negation, and explicitly annotated the negation scope (Section 5.4.4).

For the purpose of the annotation, we created a web-based annotation tool, Pair-Anno, capable of annotating aligned pairs of discontinuous scopes in two different texts[5]. As the scope of each atomic phenomena is one or more sets of tokens, prior to the annotation we automatically tokenized the corpus using NLTK [Bird et al., 2009].

5.4.1 Non-Sense Preserving Atomic Phenomena

Textual non-paraphrases in the MRPC corpus typically have a very high degree of lexical overlap and a similar syntactic and discourse structure. Normally, they differ only by a few elements (morphological, lexical, or structural), but the modification of these few elements leads to a substantial difference in the meaning of the two texts as a whole. The annotation of non-sense preserving phenomena aims to identify these key elements and study the linguistic nature of the modification.

When annotating atomic phenomena, our experts identified and annotated the type, the scope, and in some paraphrase types, the key element. Both the scope and the key are kept as a 0-indexed list of tokens. Examples 12a and 12b show a textual pair, annotated as non-paraphrase in the MRPC corpus. Table 5.2 shows the annotation of non-sense preserving atomic phenomena in 12a and 12b. The key differences are *"opposite polarity substitution (habitual)"* (type id 10) of "slip" with "rise", and the *"same polarity substitution (named entity)"* (type id 7) of "Friday" with "Thursday".

12a The loonie , meanwhile , continued to slip in early trading Friday .

12b The loonie , meanwhile , was on the rise again early Thursday .

[5] Screenshots of Pair-Anno can be seen at `https://github.com/venelink/ETPC`.