ETPC - A Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation

Venelin Kovatchev, M. Antònia Martí, Maria Salamó

Universitat de Barcelona

Gran Vía de les Corts Catalanes, 585, 08007 Barcelona, Spain {vkovatchev, amarti, maria.salamo}@ub.edu

Abstract

We present the Extended Paraphrase Typology (EPT) and the Extended Typology Paraphrase Corpus (ETPC). The EPT typology addresses several practical limitations of existing paraphrase typologies: it is the first typology that copes with the non-paraphrase pairs in the paraphrase identification corpora and distinguishes between contextual and habitual paraphrase types. ETPC is the largest corpus to date annotated with atomic paraphrase types. It is the first corpus with detailed annotation of both the paraphrase and the non-paraphrase pairs and the first corpus annotated with paraphrase and negation. Both new resources contribute to better understanding the paraphrase phenomenon, and allow for studying the relationship between paraphrasing and negation. To the developers of Paraphrase Identification systems ETPC corpus offers better means for evaluation and error analysis. Furthermore, the EPT typology and ETPC corpus emphasize the relationship with other areas of NLP such as Semantic Similarity, Textual Entailment, Summarization and Simplification.

Keywords: Paraphrasing, Paraphrase Typology, Paraphrase Identification

1. Introduction

The task of Paraphrase Identification (PI) consists of comparing two texts of arbitrary size in order to determine whether they have approximately the same meaning. The most common approach to PI is as a binary classification problem, in which a system learns to make correct binary predictions (paraphrase or non-paraphrase) for a given pair of texts. The task of PI is challenging from more than one point of view. From the resource point of view, defining the task and preparing high quality corpora is a nontrivial problem due to the complex nature of "paraphrasing". From the application point of view, for a system to perform well on PI often requires a complex ML architecture and/or a large set of manually engineered features. From the evaluation point of view, the classical task of PI does not offer many possibilities for detailed error analysis, which in turn limits the reusability and the improvement of PI systems.

In the last few years, researchers in the field of paraphrasing have adopted the approach of decomposing the meta phenomenon of "*textual paraphrasing*" into a set of "*atomic paraphrase*" phenomena, which are more strictly defined and easier to work with. "*Atomic paraphrases*" are hierarchically organized into a typology, which provides a better means to study and understand paraphrasing. While the theoretical advantages of these approaches are clear, their practical implications have not been fully explored. The existing corpora annotated with paraphrase typology are limited in size, coverage and overall quality. The only corpus of sufficient size to date annotated with paraphrase typology is the corpus by Vila et al. (2015), which contains 3900 re-annotated "*textual paraphrase*" pairs from the MRPC corpus (Dolan et al., 2004).

The use of a paraphrase typology in practical tasks has several advantages. First, "*atomic paraphrases*" are much more strict in their definition, which makes the results more useful and understandable. Second, the more detailed annotation can be useful to (re)balance binary PI corpora in terms of type distribution. Third, annotating a corpus with paraphrase types provides much better feedback to the PI systems and allows for a detailed, per-type error analysis. Fourth, enriching the corpus and improving the evaluation can provide a linguistic insight into the workings of complex machine learning systems (i.e. Deep Learning) that are traditionally hard to interpret. Fifth, corpora annotated with a paraphrase typology open the way for new research and new tasks, such as "PI by type" or "Atomic PI in context". Finally, decomposing "*textual paraphrases*" can help relate the task of PI to tasks such as Recognizing Textual Entailment, Text Summarization, Text Simplification, and Question Answering.

In this paper, we present the Extended Typology Paraphrase Corpus (ETPC), the result of annotating the MRPC (Dolan et al., 2004) corpus with our Extended Paraphrase Typology (EPT). EPT is oriented towards practical applications and takes inspiration from several authors that work on the typology of paraphrasing and textual entailment. To the best of our knowledge, this is the first attempt to make a detailed annotation of the linguistic phenomena involved in both the positive (paraphrases) and negative (nonparaphrases) examples in the MRPC (for a total size of 5801 textual pairs). The focus on non-paraphrases and the qualitative and quantitative comparison between "*textual paraphrases*" and "*textual non-paraphrases*" provides a different perspective on the PI task and corpora.

As a separate layer of annotation, we have identified all pairs of texts that include negation and we have annotated the negation scope. This makes ETPC the first corpus that is annotated both with paraphrasing and with negation.

The rest of this article is organized as follows. Section 2. is devoted to the Related Work. Section 3. describes the proposed Extended Typology, the reasons and the practical considerations behind it. Section 4. explains the annotation process, the annotation scheme and instructions, the tool that we used and the corpus preprocessing. Section 5.

presents ETPC, with its structure and type distribution. It discusses the results of the annotation and outlines some of the practical applications of the corpus. Finally, Section 6. concludes the article and outlines the future work.

2. Related Work

The task of PI is one of the classical tasks in NLP. Several corpora can be used in the task for training and/or for evaluation. Traditionally, PI is addressed using the MRPC corpus (Dolan et al., 2004). The MRPC corpus consists of 5801 pairs, that have been manually annotated as paraphrases or non-paraphrases. More recently, Ganitkevitch et al. (2013) introduce PPDB - a very large automatic collection of paraphrases, which consists of 220 million pairs. The introduction of PPDB allowed for the training of deep learning systems, due to the significant increase of the available data. However, the quality of the PPDB pairs is much lower than those of MRPC, which makes it less reliable for evaluation. A common approach is to work on both datasets simultaneously - using the PPDB for training, and the MRPC for development and evaluation.

Closely related to the PI task is the yearly task of Recognizing Textual Entailment (RTE) (Dagan et al., 2006), which has also produced various datasets and multiple practical systems. The meta-phenomena of paraphrasing and textual entailment are very similar and are often studied together at least from a theoretical point of view. Androutsopoulos and Malakasiotis (2010) present a summary of the tasks related to both paraphrasing and textual entailment.

The idea of decomposing paraphrasing into simpler and easier to define phenomena has been growing in popularity in the last few years. Bhagat (2009) and later Bhagat and Hovy (2013) propose a simplified framework that identifies several possible phenomena involved in the paraphrasing relation. Vila et al. (2014) propose a more complex, hierarchically structured typology that studies the different phenomena at the corresponding linguistic levels (lexical, morphological, syntactic, and discourse). More recently, Benikova and Zesch (2017) approach the problem by focusing on the paraphrasing at the level of events, understood as predicate-argument structure.

A similar decomposition tendency is noticed in the field of Textual Entailment. Garoufi (2007), Sammons et al. (2010), and Cabrio and Magnini (2014) propose different frameworks for decomposing the textual "inference" into simple, atomic phenomena. It is important to note that the similarity and the relation between paraphrasing and textual entailment is even stronger in the context of the decomposition framework and the resulting typologies. The two most exhaustive typologies: Vila et al. (2014) for paraphrasing and Cabrio and Magnini (2014) for textual entailment share the majority of their atomic phenomena as well as the overall structure and organization of the typology.

One of the advantages of the decomposition approaches is that naturally they work towards bridging the gap between the research at different granularity levels. A corpora annotated with semantic relations at both the textual and the atomic (morphological, lexical, syntactic, discourse) levels can be a valuable resource for studying the relation between them. In this same line of work, Shwartz and Dagan (2016) emphasize the importance of studying lexical entailment "in context" and the lack of resources that can enable such work. The corpora annotated with atomic paraphrase and atomic entailment phenomena can be used for that purpose without adaptation or additional annotation.

The application of paraphrase typology for the creation of resources and in practical tasks is still very limited. Most of the authors annotate a very small subsamples of around 100 text pairs to illustrate the proposed typology. The largest available corpus annotated with paraphrase types to date is the one of Vila et al. (2015). Barrón-Cedeño et al. (2013) use this corpus to demonstrate some possible uses of the decomposition approach to paraphrasing.

3. Extended Paraphrase Typology

We propose the Extended Paraphrase Typology (EPT), which was created to address several of the practical limitations of the existing typologies and to provide better resources to the NLP community. EPT ha better coverage than previous typologies, including the annotation of nonparaphrases. This allows for a more in-depth understanding of the meta-phenomena and of the relation between "*textual paraphrases*" and "*atomic paraphrases*".

3.1. Basic Terminology

In order to discuss the issues and limitations of existing paraphrase typologies, we first define "*paraphrasing*", "*textual paraphrase*", and "*atomic paraphrase*".

We understand "*paraphrasing*" to be a specific semantic relation between two texts of arbitrary length. The two texts that are connected by a paraphrase relation have approximately the same meaning. We call them "*textual paraphrases*". There is no limitation for "*textual paraphrases*" in terms of the nature of the linguistic phenomena involved. The concept of "*textual paraphrases*" is a practical simplification of a complex linguistic phenomenon, which is adopted in most paraphrase-related tasks, datasets, and applications. The original annotation of the MRPC and the PPDB corpora is built around the notion of textual paraphrases. Another term that we use in the article is "*textual non-paraphrases*". With this term we refer to pairs of texts (of arbitrary length), which are not connected by a paraphrase relation.

"Atomic paraphrases" are paraphrases of a particular type. They must satisfy specific (linguistic) conditions, defined in the paraphrase typology. "Atomic paraphrases" are identified by the linguistic phenomenon which is responsible for the preservation of the meaning between the two texts. "Atomic paraphrases" have a (linguistically defined) scope, such as a word, a phrase, an event, or a discourse structure. The most complete typologies to date organize "atomic paraphrases" hierarchically, in terms of the linguistic level of the involved phenomenon. Unlike "textual paraphrases", "atomic paraphrases" cannot be of arbitrary length. Their length is defined and restricted by their scope.

3.2. From Atomic to Textual Paraphrases

The relation between textual and atomic paraphrases is not easy to define and explore. It poses many challenges to the researchers, annotators, and developers of practical systems. In this section, we illustrate several issues that we want to address with the creation of the EPT and the ETPC. The first issue to be addressed is that multiple atomic paraphrases can appear in a single textual paraphrase pair. The two texts in 1a and 1b are textual paraphrases¹. However, they include more than one atomic paraphrase': *"magistrate"* and *"judge"* are an instance of *"same polarity substitution"*, while *"A federal magistrate ... ordered"* and *"Zuccarini was ordered by a federal judge..."* are an instance of *"diathesis alternation"*².

- 1a A federal **magistrate** in Fort Lauderdale ordered him held without bail.
- 1b <u>Zuccarini was ordered</u> held without bail Wednesday by a federal **judge** in Fort Lauderdale, Fla.

Second issue is that atomic paraphrases can appear in textual pairs that are not paraphrases. The two texts in 2a and 2b as a whole are not textual paraphrases, even if they have a high degree of lexical overlap and a similar syntactic structure. However, "*Microsoft*" and "*shares of Microsoft*" are an instance of "*same polarity substitution*" both phrases have the same role and meaning in the context of the two sentences. This demonstrates the possibility of atomic paraphrases being present in textual nonparaphrases. ³

- 2a <u>Microsoft</u> fell 5 percent before the open to \$27.45 from Thursday's close of \$28.91.
- 2b <u>Shares in Microsoft</u> slipped 4.7 percent in after-hours trade to \$27.54 from a Nasdaq close of \$28.91.

Third issue is that in certain cases, the semantic relation between the elements in an atomic paraphrase can only be interpreted within the context (as shown in the work of Shwartz and Dagan (2016)). The two texts in 3a and 3b are textual paraphrases. The out-of-context meaning of "cargo" and "explosives" differs significantly, however within the given context, they are an instance of "same polarity substitution".

- 3a They had published an advertisement on the Internet on June 10, offering the cargo for sale, he added.
- 3b On June 10, the ship's owners had published an advertisement on the Internet, offering the explosives for sale.

And finally, 4a and 4b illustrate an issue that is often overlooked in theoretical paraphrase research: the linguistic phenomena behind certain atomic paraphrases do not always preserve the meaning. The meanings of "*beat*" and "*battled*" are similar, and play the same syntactic and discourse role in the structure of the texts. Therefore, the substitution of "*beat*" for "*battled*" fulfills the formal requirements of a "*same polarity substitution*". However, after this substitution, the resulting texts are not paraphrases as they differ substantially in meaning.

- 4a He <u>beat</u> testicular cancer that had spread to his lungs and brain.
- 4b Armstrong, 31, <u>battled</u> testicular cancer that spread to his brain.

3.3. Objectives of EPT and Research Questions.

We argue that the objectives behind a paraphrase typology are twofold: 1) to classify and describe the linguistic phenomena involved in paraphrasing (at the atomic level); and 2) to provide the means to study the function of atomic paraphrases within pairs of texts of arbitrary size and with various semantic relations (such as, textual paraphrases, textual entailment pairs, contradictions, and unrelated texts).

Traditionally, the authors of paraphrase typologies have focused on the first objective while the latter is mentioned only briefly or ignored altogether. In our work, we want to extend the existing work on paraphrase typology in the direction of Objective 2, as we argue that it is crucial for applications. We pose four research questions, that we aim to address with the creation of EPT and ETPC:

- RQ1 what is the relation between atomic and textual paraphrases considering the distribution of atomic paraphrases in textual paraphrases?
- RQ2 what is the relation between atomic paraphrases and textual non-paraphrases considering the distribution of atomic paraphrases in textual non-paraphrases?
- RQ3 what is the role of the context in atomic paraphrases?
- RQ4 in which cases do the linguistic phenomena behind an atomic paraphrase preserve the meaning and in which they do not?

3.4. The Extended Paraphrase Typology

The full Extended Paraphrase Typology is shown in Table 1. It is organized in seven meta categories: "Morphology", "Lexicon", "Lexico-syntax", "Syntax", "Discourse", "Other", and "Extremes". Sense Preserving (Sens Pres.) shows 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.

¹All examples in this subsection are from the MRPC corpus. When we say that the texts are textual paraphrases or textual non-paraphrases, we refer to the labels corresponding to these pairs in MRPC.

²These types and annotation are from Vila et al. (2015).

³In fact, it is possible to find atomic paraphrases within pairs of texts connected by various relations, such as entailment, simplification, summarization, contradiction, and question-answering, among others. This is illustrated by the significant overlap of atomic types in Paraphrase Typology research and typology research in Textual Entailment.

⁴A more detailed table of EPT, with additional examples for each atomic type is available at https://github.com/venelink/ETPC

ID	Туре	Sense Pres.			
	Morphology-based changes				
1	Inflectional changes	+/-			
2	Modal verb changes	+			
3	Derivational changes	+			
	Lexicon-based changes				
4	Spelling changes	+			
5	Same polarity substitution (habitual)	+			
6	Same polarity substitution (contextual)	+/-			
7	Same polarity sub. (named entity)	+/-			
8	Change of format	+			
	Lexico-syntactic based changes				
9	Opposite polarity sub. (habitual)	+/-			
10	Opposite polarity sub. (contextual)	+/-			
11	Synthetic/analytic substitution	+			
12	Converse substitution	+/-			
	Syntax-based changes	•			
13	Diathesis alternation	+/-			
14	Negation switching	+/-			
15	Ellipsis	+			
16	Coordination changes	+			
17	Subordination and nesting changes	+			
	Discourse-based changes				
18	Punctuation changes	+			
19	Direct/indirect style alternations	+/-			
20	Sentence modality changes	+			
21	Syntax/discourse structure changes	+			
Other changes					
22	Addition/Deletion	+/-			
23	Change of order	+			
24	Semantic (General Inferences)	+/-			
Extremes					
25	Identity	+			
26	Non-Paraphrase	-			
27	Entailment	-			

Table 1: Extended Paraphrase Typology

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 (con*textual*)")

- 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 <u>approached</u> 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 <u>was</u> 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 <u>Mr. Hussein</u>'s daughters had British schools and hospitals in mind when they decided to ask for asylum.
- 7b "<u>Saddam</u>'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.
- 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 sensepreserving.

- 10a It was with difficulty that the course of <u>streets</u> 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 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 ("nonparaphrases") 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.

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 4.1.) in the textual non-paraphrases. Second, we annotated the sense preserving atomic paraphrase phenomena (Section 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 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 texts⁵. 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).

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 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 .

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

Table 2: Non-sense preserving phenomena

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 English⁶.

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.

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 3

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

⁶The full annotation guidelines for both sense preserving and non-sense preserving phenomena can be found at https://github.com/venelink/ETPC

- 15a Amrozi accused his brother, whom he called " the witness ", of deliberately distorting his evidence.
- 15b Referring to him as only " the witness ", Amrozi accused his brother of deliberately distorting his evidence .

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

Table 3: Sense preserving phenomenon

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.

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 nonparaphrase 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 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.

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

Table 4: Inter-annotator Agreement

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 TPOpartial (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.

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. The ETPC corpus

This section presents the results of the annotation of the ETPC corpus. Section 5.1. shows the results of annotating non-sense preserving phenomena. Section 5.2. shows the

results of annotating sense preserving phenomena. Section 5.3. discusses the results and the Research Questions, and Section 5.4. lists some applications of ETPC.

5.1. Non-Sense Preserving Atomic Phenomena

Table 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 found⁷. The total number of non-sense preserving phenomena is 3406 in 1901 text pairs.

Туре	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%

Table 5: Distribution of non-sense preserving phenomena

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.

5.2. Sense Preserving Atomic Phenomena

Table 6 shows the distribution of sense preserving atomic phenomena in the textual paraphrase and non-paraphrase portions of the corpus⁸. 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".

Туре	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%

 Table 6: Distribution of Sense preserving phenomena in

 textual paraphrases and textual non-paraphrases

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.

5.3. Discussion

In this section we briefly discuss the annotation results and the Research Questions that we posed in Section 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-paraphrase portions of

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

⁸At 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

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 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.

5.4. 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 information, 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 5.3. 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

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

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 XML⁹.

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 "textual" 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.

7. Acknowledgments

We would like to thank dr. Horacio Rodriguez and dr. Irina Temnikova for their support and suggestions, Montse Nofre and Marina Bolea for their help with the annotation, and the anonymous reviewers for their feedback and comments. This work has been funded by Spanish Ministery of Economy Project TIN2015-71147-C2-2, by the CLiC research group (2017 SGR 341), and by the APIF grant of the first author.

⁹We 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).

8. Bibliographical References

- Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. J. Artif. Int. Res., 38(1):135–187, May.
- Barrón-Cedeño, A., Vila, M., Martí, M. A., and Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947.
- Benikova, D. and Zesch, T. (2017). Same same, but different: Compositionality of paraphrase granularity levels. In *Proceedings of RANLP 2017*.
- Bhagat, R. and Hovy, E. H. (2013). What is a paraphrase? *Computational Linguistics*, 39(3):463–472.
- Bhagat, R. (2009). *Learning Paraphrases from Text*. Ph.D. thesis, Los Angeles, CA, USA. AAI3368694.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.
- Cabrio, E. and Magnini, B. (2014). Decomposing semantic inferences.
- Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.
- Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004*, pages 350–356, Geneva, Switzerland, Aug 23–Aug 27. COLING.
- Ganitkevitch, J., Durme, B. V., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In Lucy Vanderwende, et al., editors, *HLT-NAACL*, pages 758–764. The Association for Computational Linguistics.
- Garoufi, K. (2007). Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. Master's thesis, Saarland University, September.
- Sammons, M., Vydiswaran, V. G. V., and Roth, D. (2010). "ask not what textual entailment can do for you...". In ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, pages 1199–1208.
- Shwartz, V. and Dagan, I. (2016). Adding context to semantic data-driven paraphrasing. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 108–113, Berlin, Germany, August. Association for Computational Linguistics.
- Vila, M., Martí, M. A., and Rodríguez, H. (2014). "is this a paraphrase? what kind? paraphrase boundaries and typology.". pages 205–218.
- Vila, M., Bertran, M., Martí, M. A., and Rodríguez, H. (2015). Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation*, 49(1):77– 105.