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 (non-paraphrases) 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 5.2 is devoted to the Related Work. Section 5.3 describes the proposed Extended Typology, the reasons and the practical considerations behind it. Section 5.4 explains the annotation process, the annotation scheme and instructions, the tool that we used and the corpus preprocessing. Section 5.5 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 5.6 concludes the article and outlines the future work.

5.2 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.

5.3 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 non-paraphrases. This allows for a more in-depth understanding of the meta-phenomena and of the relation between *"textual paraphrases"* and *"atomic paraphrases"*.

5.3.1 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.

5.3.2 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 paraphrases[1]. However, they include more than one atomic paraphrase': *"magistrate"*

[1] All 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