The rest of this article is organized as follows. Section 8.2 lists the Related Work. Section 8.3 presents the typology, the objectives behind it and the process of selection of the types. Section 8.4 describes the annotation process - the corpus, the annotation guidelines, and the annotation interface. Section 8.5 shows the results of the annotation. Section 8.6 discusses the implications of the findings and the way our results relate to our objectives and research questions. Finally, Section 8.7 concludes the paper and addresses the future work.

8.2 Related Work

The last several years have seen an increasing interest towards the decomposition of paraphrasing [Bhagat and Hovy, 2013, Vila et al., 2014, Benikova and Zesch, 2017, Kovatchev et al., 2018a], textual entailment [Sammons et al., 2010, LoBue and Yates, 2011, Cabrio and Magnini, 2014], and textual similarity [Agirre et al., 2016].

Sammons et al. [2010] argue that in order to process a complex meaning relation such as textual entailment a competent speaker has to take several "inference steps". This means that a meta-relation such as paraphrasing, textual entailment, or semantic similarity can be "decomposed" or broken down into such "inference steps". These "inference steps", traditionally called "types" can be either linguistic or reason-based in their nature. The linguistic types require certain linguistic capabilities from the speaker, while the reason-based types require common-sense reasoning and world knowledge.

The different authors working on decomposing meaning relations all follow a similar approach. First, they propose a typology - a set of "atomic" linguistic and/or reasoning types involved in the inference process of the particular meta-relation (paraphrasing, entailment, or similarity), Then, they use the "atomic" types in a corpus annotation and finally, they analyze the distribution and correlation of the types. The corpus based studies have demonstrated that different atomic types can be found in various corpora for paraphrasing, textual entailment, and semantic similarity research.

Kovatchev et al. [2019b] empirically demonstrated that the performance of a Paraphrase Identification (PI) system on each candidate-paraphrase pair depends on the "atomic types" involved in that pair. That is, they showed that state-of-the-art automatic PI systems process "atomic paraphrases" in a different manner and with a statistically significant difference in quantitative performance (Accuracy and F1). They show that more frequent and relatively simple types like "lexical substitution", "punctuation changes" and "modal verb changes" are easier across multiple automated PI systems, while other types like "negation switching", "ellipsis" and "named entity reasoning" are much more challenging.

Similar observations have been made in the field of Textual Entailment. Gururangan et al. [2018] discovered the presence of annotation artifacts that enable models that take into account only one of the texts (the hypothesis) to achieve performance substantially higher than the majority baselines in SNLI and MNLI. Glockner et al. [2018] showed that models trained with SNLI fail to resolve new pairs that require simple lexical substitution. Naik et al. [2018] create label-preserving adversarial examples and conclude that automated NLI models are not robust. Wallace et al. [2019] introduce universal triggers, that is, sequences of tokens that fool models when concatenated to any input. All these authors identify different problems and biases in the datasets and the systems trained on them. However they focus on a single phenomenon and/or a specific linguistic construction. A typology-based approach can evaluate the performance and robustness of automated systems on a large variety of tasks.

One limitation of the different decompositional approaches is that there exist many different typologies and each typology is created considering only one meaning relation (paraphrasing, textual entailment, textual similarity). This follows the traditional approach in the research on meaning relations: each relation is studied in isolation, with its own theoretical concepts, datasets, and practical tasks.

In recent years, the "single relation" approach has been questioned by several authors. Androutsopoulos and Malakasiotis [2010] analyze the relations between paraphrasing and textual entailment. Marelli et al. [2014] present SICK: a corpus that studies entailment, contradiction, and semantic similarity. Lan and Xu [2018a] and Aldarmaki and Diab [2018] explore the transfer learning capabilities between paraphrasing and textual entailment. Gold et al. [2019] present a corpus that is annotated for paraphrasing, textual entailment, contradiction, specificity, and textual similarity. These works demonstrate that the different meaning relations can be studied together and can benefit from one another.

However, to date, the joint research of meaning relations is limited only to the binary textual labels. There has been no work on comparing the different typologies and the way different relations can be decomposed. None of the existing typologies is fully compatible with multiple meaning relations, which further restricts the research in this area. We aim to address this research gap in this paper.

8.3 Shared Typology for Meaning Relations

This section is organized as follows. Section 8.3.1 presents the problem of decomposing meaning relations. Section 8.3.2 describes our proposed typology and the rationale behind it. Section 8.3.3 formulates our research questions.

8.3.1 Decomposing Meaning Relations

The goal behind the **S**ingle **H**uman-Interpretable Typology for **A**nnotating Meaning **Rel**ations (SHARel) is to come up with a unified list of linguistic and reason-based phenomena that are required in order to determine the meaning relations that hold between two texts. The list of types should not be limited to texts that hold a specific single textual relation, such as paraphrasing, textual entailment, contradiction, and textual specificity. Rather, the types should be applicable to texts holding multiple different relations.

7 **a** All children *receive* the same education.
b The same education *is received* by all kids.

8 **a** All children *receive* the same education.
b The same education *is* **not** *received* by all kids.

In 7a and 7b, the meaning relation at a textual level is paraphrasing, while in 8a and 8b, the textual relation is contradiction. In order to determine the meaning relation for both 7 and 8, a competent speaker or an automated system needs to make several inference steps. First, they have to determine that "kids" and "children" have the same meaning and the same syntactic and semantic role in the texts. Second, they need to account for the change in grammatical voice. In terms of typology, these inference steps involve two different types - "same polarity substitution" ("kids" - "children") and "diathesis alternation" ("receive" - "is received"). In addition, in example 8b, the human or the automated system needs to determine the presence and the function of "negation" (**not**).

By successfully performing all necessary inference steps, the human (or the automated system) is able to determine that in the pair 7a-7b there is equivalence of the expressed meaning, while in the pair 8a-8b there is a logical contradiction. The required inference steps in the two examples are not specific to the textual label (paraphrasing or contradiction). The "types" are general linguistic or reason-based phenomena.

With the goal of addressing such situations, we propose a list of types that, following the existing theoretical research, can be applied to multiple meaning relations. We justify the choice of types for SHARel in the context of existing typologies.

8.3.2 The SHARel Typology

Table 8.1 shows the SHARel Typology and its 34 different types, organized in 8 categories. The first 6 categories (morphology, lexicon, lexico-syntactic, syntax,