have the same label, the same degree of complexity, and the same weight in the final evaluation of the system. However, when looking at the examples, the human intuition would suggest that:

- Processing Examples 4 and 5 requires different (linguistic) capabilities and follows different (linguistic) strategies.
- Example 5 is arguably harder than Example 4.

These intuitions contradict the empirical assumptions concerning the atomic nature of the data. If the meaning relations are indeed atomic and non-decomposable, then Examples 4 and 5 should have approximately the same degree of complexity and determining the correct label should require similar linguistic capacities and strategies.

Starting from linguistic theory and from examples like 4 and 5, several researchers have questioned the atomic nature of the Paraphrasing, Textual Entailment, and Semantic Similarity meaning relations. The "non-atomic" approach of studying meaning relations historically began in the field of Textual Entailment with the works of Garoufi [2007] and Sammons et al. [2010]. Later on Cabrio and Magnini [2014] carried out a large theoretical and empirical study on the nature of the phenomena involved in entailment. Independently from the research on textual entailment, Vila et al. [2014] and Bhagat and Hovy [2013] proposed different ways to decompose and characterize the paraphrasing relation. In the area of semantic similarity, Agirre et al. [2016] proposed a new task of “interpretable semantic textual similarity”.

In the context of this thesis, there are two important hypotheses, shared by the majority of the authors working on decomposing meaning relations.

The first hypothesis argues that in order to determine the meaning relation that holds between two texts, a human or an automated system needs to make one (or more) simple “inference steps”. In Example 4, such inference steps would be:

1) determining that “kids” in Example 4.1 means the same as “children” in Example 4.2 within the given context.

2) determining that all of the linguistic units in the two sentences in Example 4 are the same, except for “kids” - “children”.

Based on 1) and 2), a human or an automated system can determine that in Example 4, the two texts have approximately the same meaning and therefore the correct label is “paraphrases”. The hypothesis argues that to correctly predict the textual meaning relation in Example 4, a human or an automated system needs to have the capabilities and the background knowledge to process each individual “inference step”.

The second hypothesis argues out that a single example can contain various numbers of "inference steps". Example 4 has two inference steps. Example 5 has one additional step: the substitution of "receive" with "is provided to" and the corresponding change in the syntactic structure of the two sentences. Following from this hypothesis, the different number and nature of inference steps would result in different strategies for processing the examples and different degrees of complexity.

All of the authors working on decomposing meaning relations propose a list of linguistic phenomena that can be considered to be inference steps. In the rest of this dissertation these lists are called "typologies". In Table 1.1 I compare the different typologies. I also include the data for the two typologies proposed in this thesis: EPT and SHARel, presented in Chapters 5 and 8.

Table 1.1 Typologies of textual meaning relations

Typology	**Relation**	**Types**	**Lvls**	**Neg-Ex**	**Corpus**
Garoufi [2007]	TE	28	Yes	Yes	500 pairs
Sammons et al. [2010]	TE, CNT	22	No	Yes	210 pairs
Cabrio and Magnini [2014]	TE, CNT	36	Yes	Yes	500 pairs
Bhagat and Hovy [2013]	PP	25	No	No	355 pairs
Vila et al. [2014]	PP	23	Yes	No	3900 pairs
Agirre et al. [2016]	STS	9	No	Yes	3000 pairs
EPT (Chapter 5)	PP	27	Yes	Yes	5801 pairs
SHARel (Chapter 8)	PP, STS, TE, CNT	34	Yes	Yes	520 pairs

Table 1.1 compares typologies of textual meaning relations in terms of:

Relation: The textual meaning relation (or relations) that can be decomposed using the typology. TE - "Textual Entailment"; CNT - "Contradiction"; PP - "Paraphrasing"; STS - "Semantic Textual Similarity".

Types: The number of phenomena in the typology.

Lvls: Whether or not the typology is organized in hierarchical levels. For example, some typologies distinguish between morphological, lexical, syntactic, etc. phenomena, while others have no explicit structure.

Neg-Ex: Whether the typology can be used to decompose and analyze negative examples (i.e.: "non-paraphrases", "non-entailment", "0 semantic similarity") or if it is only applicable to positive examples.

Corpus: The size of the available corpora annotated with the typology.

With respect to the **relation**, each typology is built around a single empirical task. The typologies of Garoufi [2007], Bhagat and Hovy [2013], Vila et al. [2014], and Agirre et al. [2016] are all built around a single textual meaning relation. The typologies of Sammons et al. [2010] and Cabrio and Magnini [2014] can be applied to two textual meaning relations: Textual Entailment and Contradiction.

Considering the number of **types**, most of the typologies contain between 23 and 28 phenomena. The majority of these phenomena are in fact shared across the typologies of paraphrasing and textual entailment. The typology for semantic textual similarity is much more simple and task specific.

Taking into account the **levels** of hierarchical structure, three of the typologies [Garoufi, 2007, Cabrio and Magnini, 2014, Vila et al., 2014] organize the types in terms of the linguistic level of the phenomena (morphological, lexical, lexico-syntactic, syntactic, discourse, reasoning). The remaining typologies propose a list of phenomena without trying to organize them.

Looking at the decomposition of **negative examples**, the typologies for textual entailment and semantic similarity can be applied to both positive and negative examples. The typologies for paraphrasing [Bhagat and Hovy, 2013, Vila et al., 2014] can only decompose pairs of text that hold a “paraphrasing” relation. They cannot be applied to “non-paraphrases”.

Finally, with respect to the size of the available **corpora**, most typologies have been used to annotate only a small corpus (200-500 text pairs). Vila et al. [2014] and Agirre et al. [2016] are the only authors that provide corpora of a size sufficient for machine learning experiments.

Table 1.1 demonstrates some clear tendencies across the different typologies. It also illustrates some important gaps in the research field. First, at the time of beginning this dissertation each of the typologies was built around a single task and focused on one (or two) textual meaning relations. There was no typology that could be applied to multiple textual meaning relations without adaptation. Second, at the time of beginning this dissertation there was no corpus of paraphrasing or textual entailment, annotated with a typology and suitable for “recognition” machine learning experiments. The corpora of Garoufi [2007], Sammons et al. [2010], Bhagat and Hovy [2013], and Cabrio and Magnini [2014] are too small in size and the corpus of Vila et al. [2014] contains only “paraphrases”, without negative examples. With the creation of EPT and SHARel, I aimed to address these gaps in the field, as shown in the last two rows of Table 1.1.

1.1.3 Joint Research on Textual Meaning Relations

Despite the obvious similarities and interactions between the textual meaning relations, the joint research on them has been very limited, both in theoretical and