in empirical aspects. Table 1.2 shows some of the most popular corpora explicitly annotated with textual meaning relations. Most of the corpora comes from the empirical tasks of PI, RTE, and STS. I also include the data for the corpus I present in Chapter 7 of this thesis.

Table 1.2 Popular corpora for textual meaning relations

Corpus	**Paraph.**	**Entailment**	**Contradiction**	**Similarity**
MRPC	Yes	No	No	No
Quora	Yes	No	No	No
Language-Net	Yes	No	No	No
RTE (1-3)	No	Yes	No	No
RTE (4-6)	No	Yes	Yes	No
SNLI	No	Yes	Yes	No
MultiNLI	No	Yes	Yes	No
STS (all)	No	No	No	Yes
SICK	No	Yes	Yes	Yes
Sukhareva et al. [2016]	Yes *	Yes	No	No
Chapter 7 (this thesis)	Yes	Yes	Yes	Yes

Table 1.2 clearly demonstrates the separation between the different meaning relations in existing corpora. Each corpus is typically built around one single relation, or two in the case of textual entailment. At the time of beginning this thesis the only corpora that contained multiple textual meaning relations were:

- the SICK corpus [Marelli et al., 2014], which is annotated for textual entailment, contradiction, and semantic similarity.
- the corpus of Sukhareva et al. [2016] who annotate paraphrasing as a specific sub-class of entailment.

The corpus presented in Chapter 7 addresses this gap in the existing resources and is the first corpus annotated with the four most popular textual meaning relations: Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity.

In a more theoretical setting, Madnani and Dorr [2010] and Androutsopoulos and Malakasiotis [2010] discuss and compare different aspects of paraphrasing and textual entailment. They argue that paraphrasing is typically a bi-directional entailment. Cabrio and Magnini [2014] and Sukhareva et al. [2016] also suggest that paraphrasing is a sub-class of textual entailment.

However, Dolan and Brockett [2005] point out that if they enforced a strict bi-directional entailment and full equivalence of the information content, the annotators would only mark identical texts as paraphrases, which would make the

Paraphrase Identification task trivial. Therefore in their annotation setup they also allow for a limited difference in the information content in the two texts. As a result, the equivalence between bi-directional entailment and paraphrasing does not hold in their corpus (MRPC). A similar approach to annotating the paraphrasing relation has also been adopted in the rest of the PI corpora. Therefore, the relation between entailment and paraphrasing is non-trivial to define in an empirical setting. However, the lack of corpora annotated for multiple textual meaning relations has limited the possibilities for empirical data-driven research on the interactions between paraphrasing, textual entailment, and contradiction.

There has also been some research on using one textual meaning relation to predict another and for the transfer of knowledge across tasks. Cer et al. [2017] argue that to find paraphrases or entailment, some level of semantic similarity must be given. Bosma and Callison-Burch [2006] use techniques from Paraphrase Identification in order to solve textual entailment. Castillo and Cardenas [2010] and Yokote et al. [2011] use semantic similarity to solve entailment.

The recent work by several authors is indicative of an increasing interest towards the joint study of meaning relations. In particular, the topic is interesting within the context of transfer learning in NLP and CL. Lan and Xu [2018a] and Aldarmaki and Diab [2018] demonstrate the transfer learning capabilities of different systems in the tasks of PI and RTE. They cover a wide range of supervised and unsupervised machine learning architectures and demonstrate promising results.

The interest and success of the transfer learning techniques have also resulted in the creation of the GLUE [Wang et al., 2018] and SuperGLUE [Wang et al., 2019] benchmarks. GLUE and Super GLUE are a collection of multiple datasets for several tasks, including PI, RTE and STS. The authors of those benchmarks argue that systems working on Natural Language Understanding (NLU) should be able to perform well on all of the tasks, and not just on one. The GLUE and SuperGLUE are now the most popular benchmarks for evaluating NLU systems and general purpose meaning representation models.

However, I would argue that a benchmark of multiple datasets is not a replacement for a single dataset annotated with multiple textual meaning relations. Similarly, a transfer learning experiment is not a replacement for a single task of multi-class classification. At the time of beginning this thesis there was an apparent gap in the field - a lack of resources (annotation guidelines and corpora) that would enable the joint theoretical and empirical research of multiple textual meaning relations.

1.1.4 Other Related Work

Distributional Semantics (DS) is the predominant framework for representing and comparing the meaning of linguistic units in contemporary Computational Linguistics (CL) and Natural Language Processing (NLP). DS has an important role both in theoretical research and in developing practical applications. The core hypothesis in DS is the Distributional Hypothesis (DH), as formulated by different authors:

> *"Difference in meaning correlated with difference in distribution"*
> [Harris, 1954]
>
> *"You shall know a word by the company it keeps"*
> [Firth, 1957]
>
> *"The meaning of a word is its use in the language"*
> [Wittgenstein, 1953]

While these authors formulate DH in slightly different ways, the central assumption remains the same and can be stated as follows: *"similar (or semantically related) linguistic units appear in similar contexts"*. This assumption allows for a radical empirical approach towards formalizing the meaning of linguistic units. There exist many Distributional Semantic Models (DSM) for representing the meaning of words or complex language expressions. Baroni and Lenci [2010], Turney and Pantel [2010], and Lapesa and Evert [2014] compare different DSMs. More recently, the popular DSMs are based on neural network architectures (Word2Vec [Mikolov et al., 2013b], Glove [Pennington et al., 2014], Skip-Thought [Kiros et al., 2015], InferSent [Conneau et al., 2017], and ELMO [Peters et al., 2018]. DSMs are used in many practical applications. They are also very popular for empirical tasks focused on textual meaning relations. Paraphrasing, Textual Entailment, and Semantic Textual Similarity are often considered evaluation benchmarks for the quality of DSMs.

Within CL and NLP there are many empirical tasks focused on meaning relations at the level of tokens (i.e.: words and multi-word expressions), henceforth **"lexical meaning relations"**. Hill et al. [2015] and Bruni et al. [2014] propose datasets for out-of-context lexical similarity, while Huang et al. [2012] and Levy et al. [2015] propose datasets for context-sensitive lexical similarity. Kremer et al. [2014] present a dataset for the "lexical substitution" task. Hendrickx et al. [2010] propose the task of "relation classification" at the lexical level.

There are many manually created resources for studying and processing lexical meaning relations. These resources include, for example, lists of words with a particular relation, morphological rules for creating a particular relation (e.g.: