The empirical tasks focused on the automatic processing of meaning relations are inspired by human capabilities. We, as competent language users, can quickly and unconsciously determine the meaning relation that holds between two simple or complex language expressions. We can successfully recognize, generate, and extract paraphrases, entailment pairs, and contradiction pairs. The empirical tasks focused on textual meaning relations in CL and NLP aim to produce automated systems that can achieve on-task performance comparable with that of humans. Human judgments are typically taken as a gold standard for evaluation.

In this thesis, I focus on recognition tasks. In particular, I study **Paraphrase Identification**, **Recognizing Textual Entailment**, and **Semantic Textual Similarity**. In the rest of this section I present the definition, corpora, and state-of-the-art for each of these three tasks.

Paraphrase Identification

Task format and definition: Paraphrase Identification (PI) is framed as a binary classification task. In PI, a human or an automated system needs to determine whether or not a paraphrasing relation holds between two given texts. The definition of "paraphrasing" provided by Dolan and Brockett [2005] is *"whether two sentences at the high level "mean the same thing" /.../ despite obvious differences in information content."*.

(1) **Sentence 1:** The genome of the fungal pathogen that causes Sudden Oak Death has been sequenced by US scientists.
Sentence 2: Researchers announced Thursday they've completed the genetic blueprint of the blight-causing culprit responsible for sudden oak death.

Two sentences that are connected with a paraphrasing relation can be seen in Example 1[3]. While the two sentences are not completely equivalent, in the context of PI they are considered paraphrases. Dolan and Brockett [2005] argue that if human annotators are required to only mark full equivalence of meaning, only identical sentences are considered paraphrases. Therefore, in the practical setting of PI, they propose a less strict definition of paraphrasing and allow for some difference in the information content.

PI Corpora: The task of (PI) was first popularized with the creation of the Microsoft Research Paraphrase Corpus (MRPC), presented in Dolan et al. [2004] and Dolan and Brockett [2005]. The MRPC corpus is semi-automatically created from the articles in the news domain and consists of 5,801 text pairs, annotated as

[3]The example is taken from Dolan and Brockett [2005].

"paraphrase" or "non-paraphrase". To date, MRPC is still used for the evaluation of automated PI systems despite, its relatively small size.

The Paraphrase Database (PPDB) [Ganitkevitch et al., 2013] (and later on its second version PPDB2 [Pavlick et al., 2015]) was the first large scale paraphrase corpus. It is an automatically constructed collection of over 100 million paraphrases at different granularity. While the MRPC only contains sentences and longer chunks of text, the PPDB also contains "paraphrases" of words and short phrases. The second version of PPDB also includes the entailment relation. PPDB and PPDB2 are collections of paraphrases, rather than corpora specifically created for the task of PI. However, they can be adapted for use in PI tasks.

The Quora Question Pair Dataset [Iyer et al., 2017] is a semi-automatically collected corpus of 400,000 question pairs marked as "duplicate" or "non-duplicate" by Quora users. The corpus was used in an online competition[4] and facilitated the use of Deep Learning based systems for the task of PI. Due to its size, the Quora corpus is very popular for training state-of-the-art PI systems.

The Language-Net corpus [Lan et al., 2017] is the largest PI dataset to date. It was extracted from Twitter and contains over 51,000 human-annotated sentence pairs and over 2.8 million automatically extracted candidate paraphrases.

MRPC, PPDB, Quora, and Language-net are all created for the English language. The work on PI for languages other than English is very limited. We can mention the work of Creutz [2018] on the creation of paraphrase corpus in six languages using open subtitles dataset.

State-of-the-art in PI: The first automated PI systems were based on manually engineered features [Finch et al., 2005, Kozareva and Montoyo, 2006] or on a combination of lexical similarity metrics and cosine similarity [Mihalcea et al., 2006]. Word2Vec [Mikolov et al., 2013b] and Glove [Pennington et al., 2014] introduced a new paradigm in PI, but also in CL and NLP in general. The systems based on Word2Vec and Glove outperformed previous unsupervised systems and pushed the state-of-the-art further. Deep Learning based systems using autoencoders [Socher et al., 2011], Long Short Term Memory Networks (LSTM) [He and Lin, 2016], and Convolutional Neural Networks (CNN) [He et al., 2015] set the new state-of-the-art for the Supervised PI systems. More recently, Transformer based architectures [Devlin et al., 2019] have made a considerable improvement to automated PI systems, approaching human level performance on the datasets[5].

[4] `https://www.kaggle.com/c/quora-question-pairs`

[5] The official ACL page for PI (`https://aclweb.org/aclwiki/Paraphrase_Identification_(State_of_the_art)`) and the GLUE benchmark page (`https://gluebenchmark.com/leaderboard`) contain the full leaderboard of PI systems for a variety of corpora.

Recognizing Textual Entailment

Task format and definition: Recognizing Textual Entailment (RTE), also known as Natural Language Inference (NLI), has two different formats. The original RTE was framed as a binary classification task. In RTE, a human or an automated system needs to determine whether or not a paraphrasing relation holds between two given texts. The practical definition of Textual Entailment in RTE is *"a directional relationship between pairs of text expressions, denoted by T - the entailing "Text", and H - the entailed "Hypothesis". We say that T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people."*. An example of textual entailment relation can be seen in 2. In the example given the Text entails Hyp 1, but not Hyp 2, or Hyp 3.

(2) **Text:** The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.
Hyp 1: BMI acquired an American company.
Hyp 2: BMI bought employee-owned LexCorp for $3.4Bn.
Hyp 3: BMI is an employee-owned concern.

The second format of the RTE was introduced in [Giampiccolo et al., 2008] and the task was reformulated as a three class classification between "entailment", "contradiction", and "neutral" text pairs. In example 2, the Text entails Hyp 1, contradicts Hyp 2, and is neutral with respect to Hyp 3.

RTE Corpora: The task of RTE was popularized with the introduction of the yearly Recognizing Textual Entailment challenge in Dagan et al. [2006].The first three editions of the RTE challenge were called the Pascal RTE challenge [Dagan et al., 2006, Bar-Haim et al., 2006, Giampiccolo et al., 2007] and were framed as a binary classification between "entailment" and "non entailment" text pairs. In the fourth edition of the challenge [Giampiccolo et al., 2008], the Pascal RTE challenge became the Text Analysis Conference (TAC) RTE challenge. The task was reformulated as a three class classification between "entailment", "contradiction", and "neutral" text pairs. The TAC RTE challenge ran for four years: Giampiccolo et al. [2008], Bentivogli et al. [2009], Bentivogli et al. [2010], and Bentivogli et al. [2011]. Like the MRPC corpus, the RTE datasets are not very large in size, however due to the high quality of the annotation they are still used as an evaluation benchmark for state-of-the-art systems.

The increasing popularity of Deep Learning systems and the need for more training data led to the creation of the Stanford Natural Language Inference corpus (SNLI) [Bowman et al., 2015] and later on the Multi-Genre Natural Language