



# Chapter 9

## Conclusions

In this final chapter, I look back at what has been done in this thesis. In Section 9.1, I summarize the main contributions of my research. I discuss the importance of my findings in the context of the research on textual meaning relations. In Section 9.2, I present a description of the different resources, created as part of this thesis and released to the community. In Section 9.3, I outline some open issues for future research on textual meaning relations and the way forward to addressing them.

### 9.1 Contributions and Discussion of the Results

The contributions of this thesis can be grouped into four thematic categories, corresponding to the four thesis objectives formulated in Section 1.2.

#### **Empirical applications of paraphrase typology**

The first **objective** of this thesis is *To use linguistic knowledge and paraphrase typology in order to improve the evaluation and interpretation of the automated paraphrase identification systems*. This objective has been addressed in the articles Kovatchev et al. [2018a], Kovatchev et al. [2018b], and Kovatchev et al. [2019b], presented in Chapters 4, 5, and 6.

Traditionally, the task of Paraphrase Identification is framed as a binary classification problem. It requires manually or semi-automatically annotated data for training and testing. The performance of the automated systems is evaluated using Accuracy and F1 measures. The state-of-the-art systems that work in Paraphrase Identification are mostly based on complex deep learning architectures and trained on large amounts of data. The linguistic intuitions and resources are relatively less important for these systems.

In this thesis I found evidence that the linguistic intuitions from the theoretical research on Paraphrase Typology can successfully be incorporated in the empirical task of Paraphrase Identification. The **contributions** of this part of the thesis are two empirical applications of paraphrase typology:

- A statistical corpus-based analysis presented in Chapter 5. I measured and compared the distribution of the different paraphrase types in the MRPC corpus. The analysis of the results shows that the type distribution is severely imbalanced: some paraphrase types appear in the majority of the examples, while other types are underrepresented. I hypothesize that this imbalance introduces a potential bias in the datasets.
- A “qualitative evaluation framework” presented in Chapter 6. I evaluated and compared the performance of 11 different automated paraphrase identification systems on each of the paraphrase types. The results depicted that Accuracy and F1 measures fail to capture important aspects of the performance of the automated systems. I showed that the performance of the evaluated systems varies significantly based on the types involved in each candidate-paraphrase pair. I also showed that systems with quantitatively similar performance can make qualitatively different predictions and errors.

This work gave a new perspective on the task of Paraphrase Identification. I argued that the “binary” definition of the task is oversimplified as it does not account for the different linguistic and reason-based phenomena involved in paraphrasing. The experiments showed that linguistic knowledge, in particular Paraphrase Typology, can be beneficial when analyzing the quality of the corpora and the performance of the automated systems.

### **In-depth knowledge about paraphrasing**

The second **objective** of this thesis is *To empirically validate and quantify the difference between the various linguistic and reason-based phenomena involved in paraphrasing*. This objective has been addressed in the articles Kovatchev et al. [2018a] and Kovatchev et al. [2019b], presented in Chapters 5 and 6.

At the beginning of this dissertation, the research on Paraphrase Typology was predominantly theoretical. The different authors proposed lists of phenomena involved in the textual meaning relations and provided examples for each different type. However, there were no empirical experiments that could measure the practical implications of the theoretical difference between the paraphrase types.

In this thesis I prepared and carried out the first empirical experiment aimed at validating the theoretical concepts of the research on Paraphrase Typology. The